MPHIL

Survival of duplicate genes in the vertebrate genome: A matter of splice or death

Xia, Bing

Award date:
2018

Awarding institution:
University of Bath

Survival of duplicate genes in the vertebrate genome: A matter of splice or death

Bing Xia

A thesis submitted for the degree of Master of Philosophy

University of Bath

Department of Biology and Biochemistry

Oct 2018

Copyright Notice

Restrictions on use and Licensing

Access to this thesis/portfolio in print or electronically is restricted until ……. (date). Signed on behalf of the Doctoral College.....(print name)……

Declaration of any previous submission of the work

The material presented here for examination for the award of a higher degree by research has / has not been incorporated into a submission for another degree.

(If applicable, provide the relevant details i.e. those parts of the work which have previously been submitted for a degree, the University to which they were submitted and the degree, if any, awarded).

Candidate's signature: 

Declaration of authorship

I am the author of this thesis, and the work described therein was carried out by myself personally, with the exception of …0……. article/chapter where …0%…… (detail the amount in percentage terms) of the work was carried out by other researchers (e.g. detail any collaborative works included in the thesis in terms of formulation of ideas, design of methodology, experimental work, and presentation of data in journal format).

Candidate's signature: 

CONTENTS

ABSTRACT

Why do some duplicate genes survive while others are lost? After duplication, most extra gene copies soon accumulate disabling mutations and degrade, but some are retained. Although there are several models proposed to explain fates of duplicated genes factors associated with duplicate gene retention remain poorly understood. Here, we analyse gene duplicate retention after whole genome duplication events at the base of the vertebrate lineage. We show that ancestral alternative splicing -a process by which a single gene can encode more than one distinct protein- in invertebrate single copy genes increased significantly the number of orthologous present in vertebrate genomes. Similar results were observed for non-singletons in the invertebrate genomes. Furthermore, we find that a higher number of orthologous in vertebrate genomes is associated with significantly lower expansion of alternative splicing consistent with the two processes being to some extent equivalent ways to increase transcript diversity. After that, bias of the reliability of our database was clarified. Finally, we found the complexity of species also provide a positive correlation with duplicates and splicing. These observations are consistent with the view that alternative splicing shapes the survival chances of duplicate genes by facilitating functional split between gene copies.

ABBREVIATION

AS: Alternative splicing

ASE: Alternative splicing events

EST: Expressed sequence tag

GFS: Gene family size

GO: Gene ontology

INTRODUCTION

Gene duplication is considered a major source of molecular and functional innovation[1-4 5-8]. However, following duplication events involving a single or a few genes or whole genome duplication events, most genes return to their original single copy status as extra copies accumulate deleterious mutations and are degraded[2]. What mechanisms determine the fate of duplicate genes has been a subject of intense speculation with numerous models proposed to explain the evolutionary paths of duplicate genes[1,4]. With the exception of genes under positive selection for increased gene dosage, most models for duplicate gene retention assume that the fate of duplicate genes is a stochastic process where gene characteristics of the single copy gene prior to the event of duplication make little difference to the fate of the duplicate copy [1,4].

Patterns of retention of duplicate genes are difficult to assess as degraded duplicate genes are hard to identify after a few million years making it hard to differentiate variations in rates of gene duplication and rates of duplicate retention. Thus, whole genome duplication events where all genes get an additional copy provide an ideal system to examine rates of duplicate retention. By examining patterns of duplicate retention in vertebrates which are thought to have undergone at least two rounds of genome duplication, Edger and colleagues[9] showed that genes locked in dosage-sensitivity interactions were retained in larger numbers after duplication as deletions affecting single genes in a dosage-sensitivity group were more disruptive.

One of the most widely accepted models driving the retention of duplicate genes involves a process of subfunctionalization with gene duplicates accumulating reciprocal translocation after which both copies are needed to perform the original function. However, little is known about factors predisposing certain genes to undergo subfunctionalisation after duplication. Alternative splicing is a process by which a single gene generates more than one protein by selectively splicing out one or more segments from its coding region. This process originally identified in the late 1970s[5,6], is now

thought to be nearly ubiquitous across eukaryotic taxa and particularly common in metazoan genomes. Alternative splicing can expand transcript diversity of genes[10-12] and has been proposed to constitute a major source of functional innovation[13,14]. Recent studies have shown that levels of alternative splicing are closely related to increased organism complexity assayed by the number of cell types[15,16].

Alternative splicing is regulated by short sequence motifs and thus a reduced number of mutations can lead to the loss and gain of alternative splicing events. Thus, alternative splicing could facilitate subfunctionalization of duplicated genes as a few mutations would suffice to lock each gene copy into producing transcripts similar to distinct alternative splicing variants produced by the ancestral non-duplicated gene. In assessing the link between gene duplicate retention and alternative splicing we can consider three scenarios. The first is that the two processes are not equivalent in their contribution to functional transcript in which case we would expect no significant association between alternative splicing and gene duplicate retention. The second scenario proposes that alternative splicing does not facilitate subfunctionalization but that it is a complementary mechanism for the regulation of transcript diversity to gene duplication. In this case, we may observe a negative or positive correlation between ancestral alternative splicing and gene duplicate retention depending on whether alternatively spliced genes have reached their optimal transcript diversity and thus there is less selection to retain duplicates or because ancestral alternative splicing is a marker of higher selection for transcript diversity, which would be likely to be accompanied by higher duplicate retention on top of existing alternative splicing levels. Finally, alternative splicing could act as an additional mechanism for regulation of transcript diversity and does facilitate subfunctionalization after a duplication event with few complementary mutations. In this case we would expect to observe higher duplicate retention of ancestrally alternatively spliced genes followed by reductions in levels of alternative splicing.

Consistent with this, a recent study examining duplicate retention after a round of genome duplication in the lineage leading to *Saccharomyces cerevisiae* found that genes with ancestral alternative splicing were more likely to be retained as duplicates compared to genes with no ancestral alternative

splicing[17]. Because *Saccharomyces cerevisiae* lost several key genes involved in the alternative splicing machinery in other eukaryotes[17], it is not possible to test the effects on alternative splicing of the retention of duplicates. In mammals, it has been proposed that, to some extent, gene duplication and alternative splicing are mutually exclusive mechanisms regulating protein diversity[18] and several studies[19-21] have found a negative correlation between alternative splicing and gene family size. Furthermore, genome-wide scale studies in mammalian species have shown that, as expected under a functional sharing model, alternative splicing is lower among newly duplicated genes[8,19]. However, this pattern has also been proposed to result from higher gene duplicate retention among genes with lower alternative splicing rates[22].

Results from analyses of single gene families offer some support for the subfunctionalisation model. Three AS isoforms of troponin I in *Ciona* have been found to have corresponding expression patterns to each of the three duplicate troponins I genes found in vertebrates[23]. Similarly, expression patterns of two AS isoforms from single copy genes in tetrapods are mirrored in duplicated synapsin-2 genes [24] and *MITF* genes in teleost fish species [25,26]. It is thus possible to hypothesise a general key role for ancestral alternative splicing in duplicate gene retention during vertebrate evolution. Two rounds of genome duplication at the base of the vertebrate lineage[3,18] provide us with ideal conditions to test this hypothesis.

Function is also been researched as factors co-related with the reservation of gene duplication. In Misook's research[27], gene duplication will facilitate the expression diversity in closely related species and allopolyploids of *Arabidopsis thaliana*. The word 'rapid' was used to describe the difference of multiple copy genes. Meanwhile the gene in charge of stress response and abiotic or biotic stimulus response showed a higher no-additive expression. In term of the complexity of the species, Nuno's study [28] revealed the alternative splicing has a positive trend along the evolutionary tree and human is in the top position.

Here, we assess the relation between ancestral alternative splicing in invertebrate species and duplicate gene retention in vertebrate species following two whole genome duplication events at the base of the vertebrate lineage. We further assess the alternative splicing status in vertebrate lineages for genes retained in duplicate form and those which returned to single copy status. If models proposing equal chances of duplicate retention are correct, then we can expect ancestral alternative splicing not to be relevant for duplicate gene retention. Alternatively, we would expect those genes with high levels of alternative splicing, to begin with, will tend to have a high rate of duplicate retention in association with lower levels of alternative splicing in the resulting multi-copy genes. Finally, we will detect the relationship between the function-domain and alternative splicing and gene family size. If duplicate and splicing have some influence on the gene function complex, we expect the different level of our dataset in different function domains.

RESULTS

Ancestral alternative splicing in invertebrate genes is associated with higher levels of gene duplicate retention in vertebrate genomes.

To investigate the relationship between alternative splicing and orthologous number, alternative splicing events (ASEs) were identified from comparing alignment with ESTs partial transcripts for three invertebrate species and six vertebrate species [29]. To correct the bias in alternative splicing event detection from differential transcript coverage among genes and species, we obtained comparable estimates of alternative splicing levels by using a transcript number normalization protocol as implemented in [30]. Orthologous number in vertebrate species for each gene in invertebrates, were obtained from Ensembl gene family annotations (Table 1).

Table 1 **Experiment Species List**

	Species Name	Amount of All Gene Family	Amount of Singletons Gene Family	Amount of ASE Data
Invertebrates	*Drosophila melanogaster*(Fruit fly)	11277	10059(89%)	4667(41%)
	Caenorthhabdits elegans(Worm)	14646	12600(86%)	2943(20%)
	Ciona intestinalis(Vase tunicate)	14166	12933(91%)	4168(29%)
Vertebrates Species	*Danio rerio*(Zebrafish)	12624	8097(64%)	7263(58%)
	Xenopus tropicalis(Xenopus)	10390	7225(70%)	5779(56%)
	Gallus gallus(Chicken)	10690	7937(74%)	4817(39%)
	Homo sapiens(Human)	13347	9369(70%)	6561(49%)
	Mus musculus(Mouse)	13376	9826(73%)	8125(61%)
	Anolis carolinensis(Anole lizard)	11180	7963(71%)	166(1.4%)

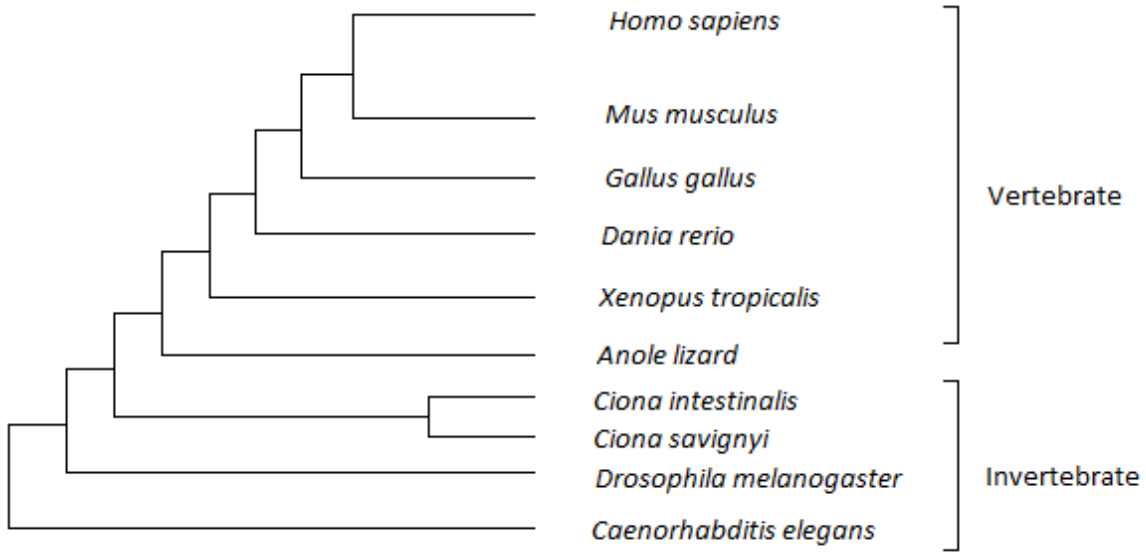


Figure 1 **Tree of evolution time of our research species.** Three species invertebrates were settled as ancestral genomic data group and the other six vertebrates experimented for our research. The tree was drawn by ourselves base on the relevant journals[31,32] *Ciona savignyi* (another species ciona) will be used in the following experiment.

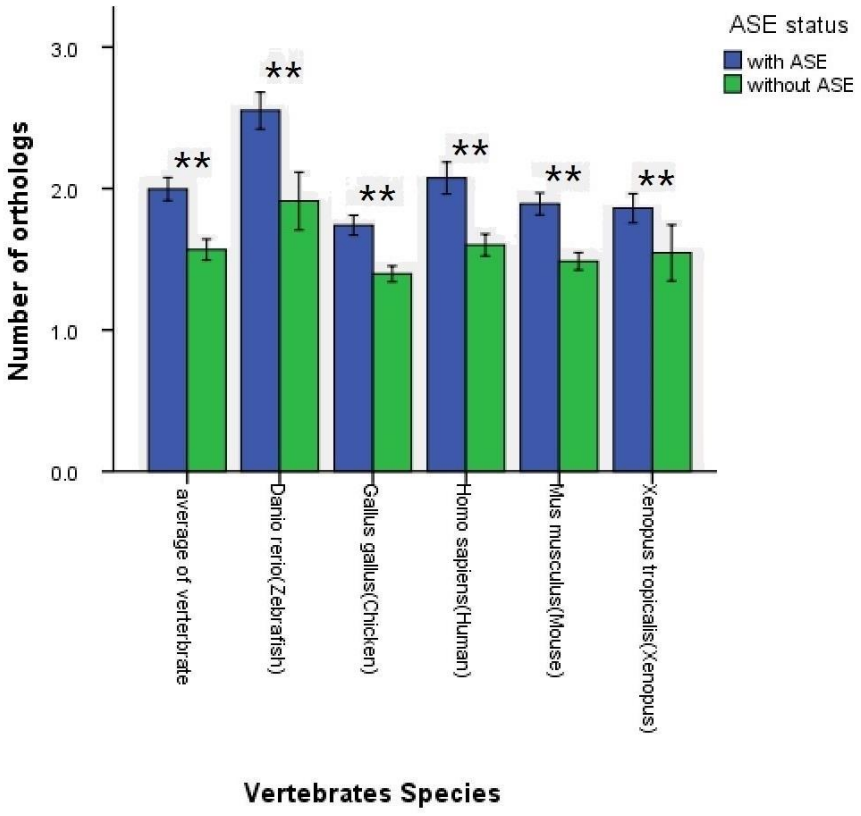
According to the Interactive tree of life and journal, we choose 9 species as our study targets. Due to the invertebrate species are in an earlier position in the life tree (Figure1) and before the 2 round WGD. We choose them as "temporary ancestor" in our research.

Table 2 **ASE status details of invertebrates**

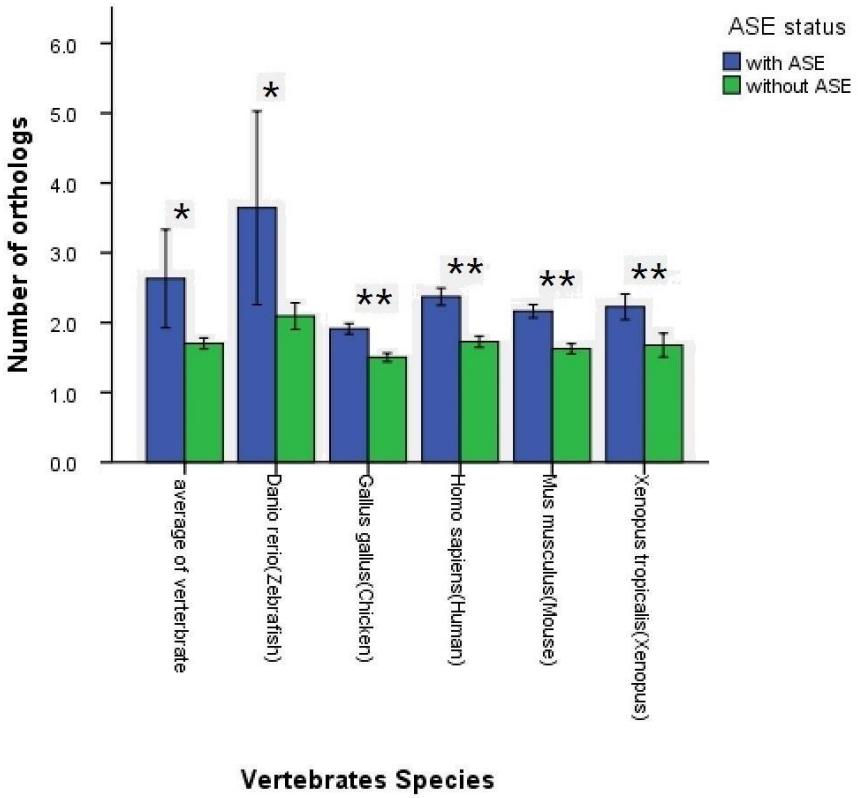
	ASE<=0.5	ASE>0.5
Fruit fly	2500	2167(46%)
Worm	1977	966(33%)
Vase tunicate	1662	2505(60%)

To begin with, we assessed the relationship between alternative splicing in genes present as single copy in the invertebrate genome of fruit fly with orthologous numbers in the vertebrate genomes for human, mouse, chicken, zebrafish and frog. When dividing these genes according to ASE in

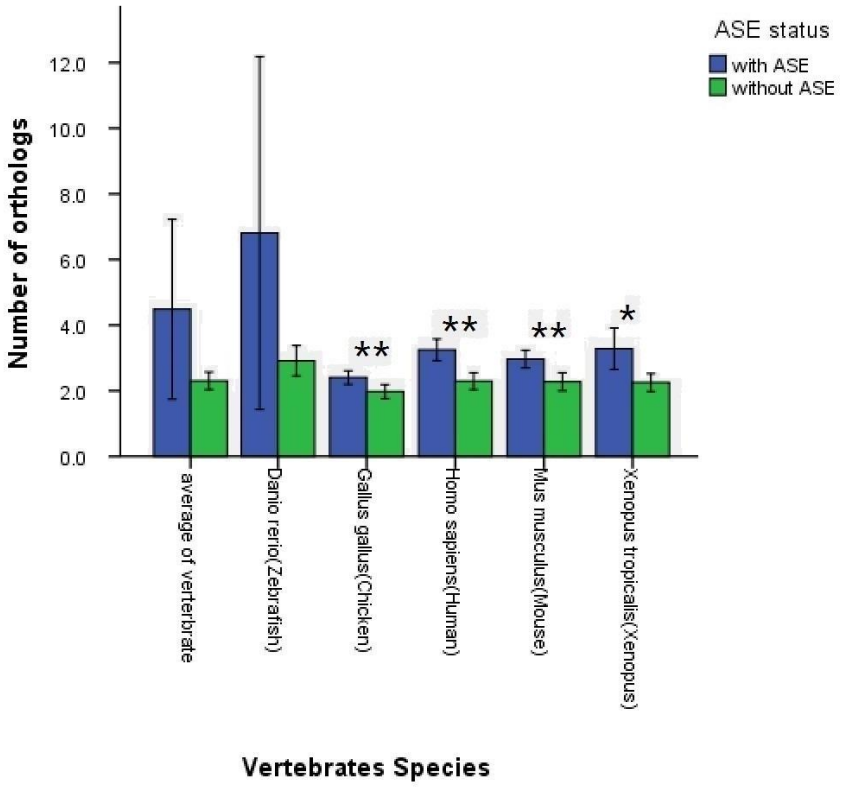
invertebrates into those without ASEs (<=0.5) and with ASE (>0.5) (Table 2) and compared the difference between the orthologous numbers of these two groups in vertebrates, we found a significant increase in orthologous number in all five vertebrate species tested (T-test; $p < 0.01$; Figure 2A). Similar patterns were found when assessing nonsingleton genes in the fruit fly genome (Figure 2BC), except for the zebra fish (+3.025, P=0.17). These findings suggest that the presence of ancestral alternative splicing of a gene in invertebrate genomes is associated with a higher number of orthologous numbers in vertebrate genomes.



A


ASE status
with ASE
without ASE
Number of orthologs
6.0
5.0
4.0
3.0
2.0
1.0
0.0
*
*
**
**
**
**
average of verterbrate
Danio rerio(Zebrafish)
Gallus gallus(Chicken)
Homo sapiens(Human)
Mus musculus(Mouse)
Xenopus tropicalis(Xenopus)
Vertebrates Species

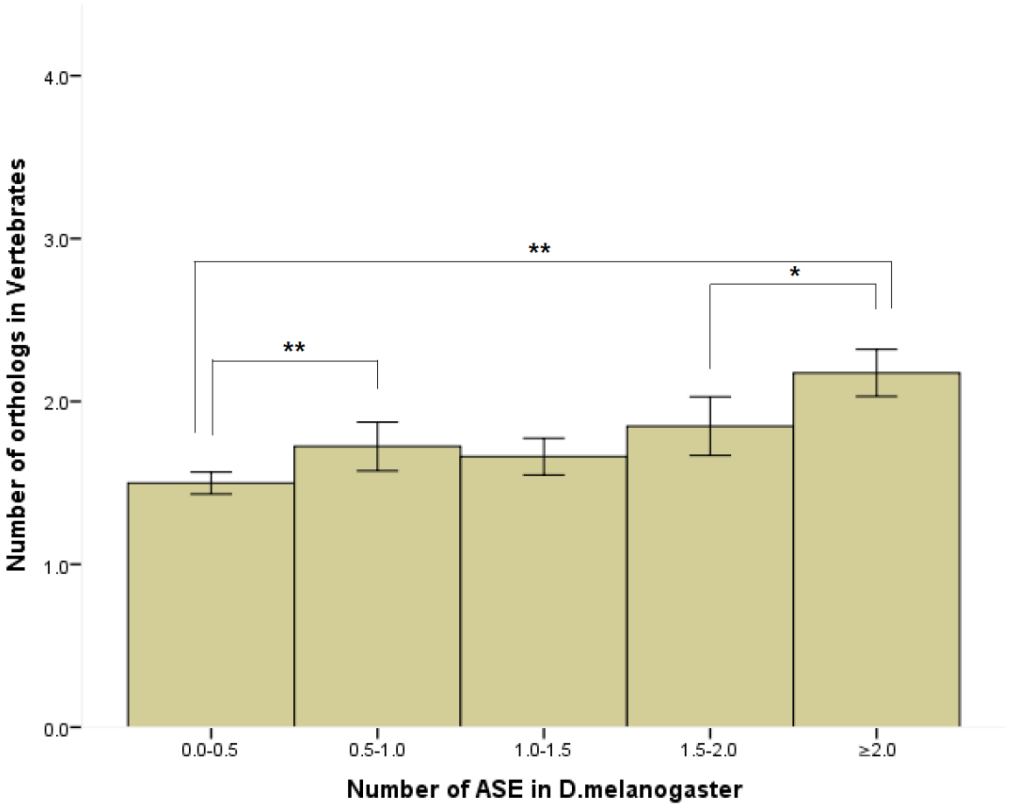
B


ASE status
with ASE
without ASE
Number of orthologs
12.0
10.0
8.0
6.0
4.0
2.0
0.0
**
**
**
*
average of verterbrate
Danio rerio(Zebrafish)
Gallus gallus(Chicken)
Homo sapiens(Human)
Mus musculus(Mouse)
Xenopus tropicalis(Xenopus)
Vertebrates Species

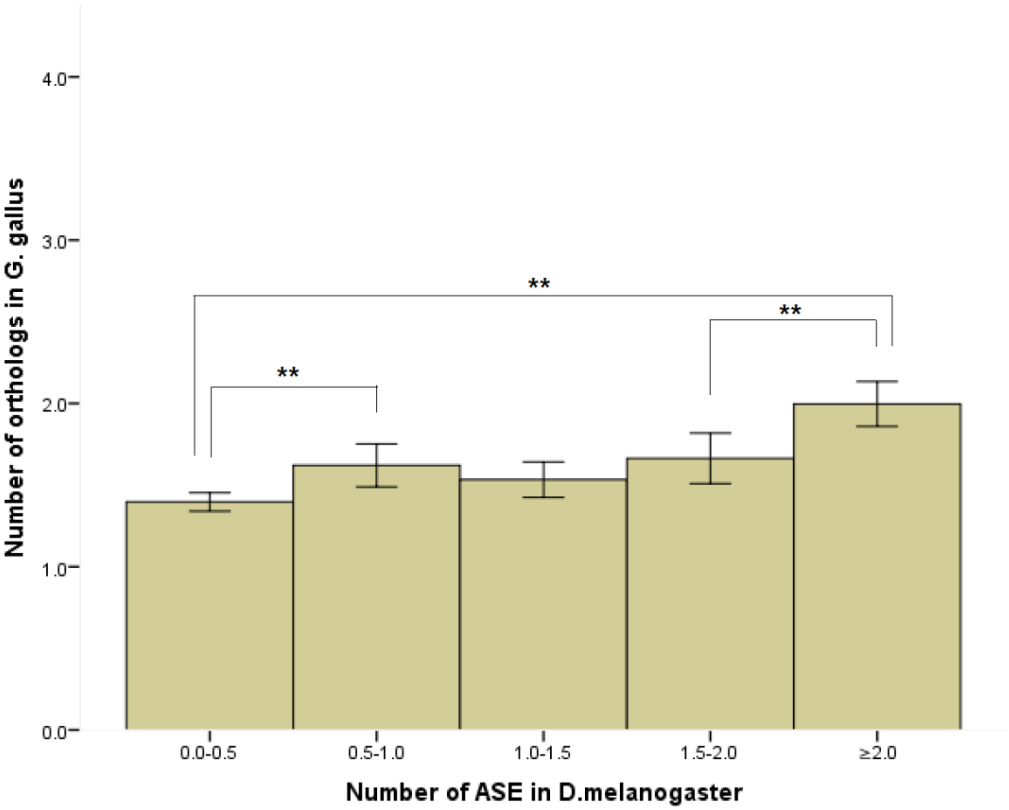
C

Figure 2 Comparison of number of orthologous between two gene categories (with and without ASE in *D. melanogaster*) in all vertebrates species (single copy (A) all genes (B) and multiple copies (C). The alternative splicing events value 0.5 was used as a line of demarcation to judge the gene with (ASE>0.5) or without (ASE<=0.5) AS. X axis was labelled by the species name. Y axis marks the mean of orthologous in vertebrates species . T-test were applied to the data set, *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.

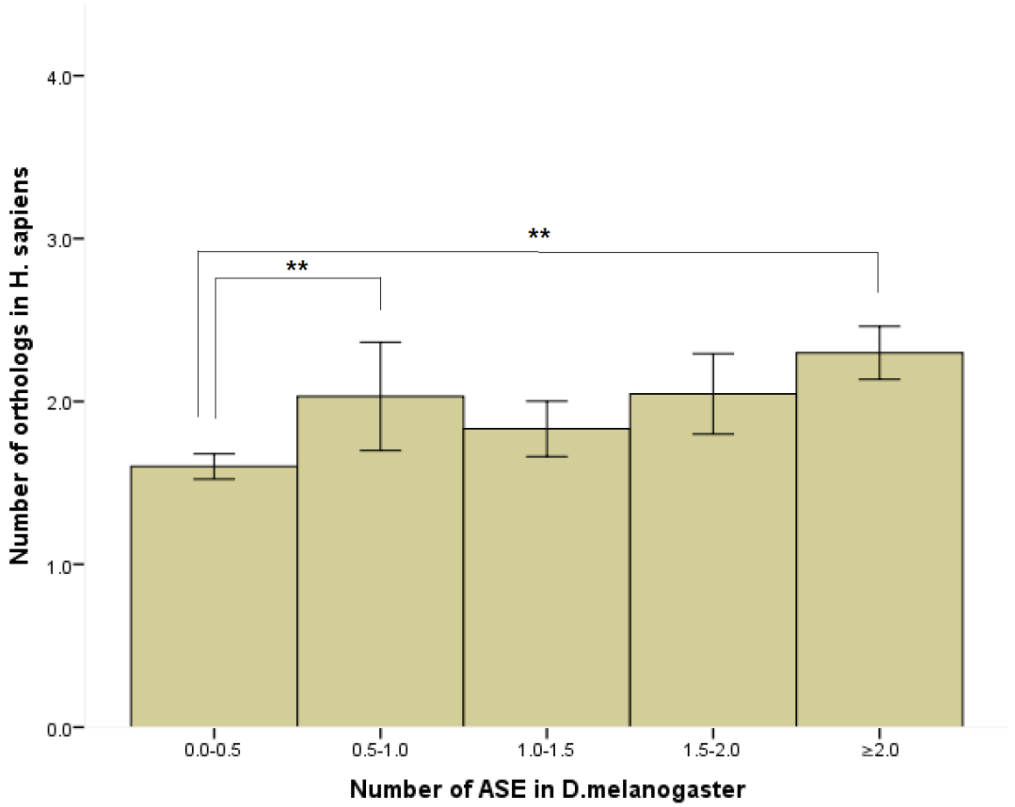
To assess if differentiation of levels of ancestral alternative splicing are associated with distinct levels of orthologous number in vertebrate genomes we divided genes into five groups based on the number of ASEs in fruit fly (0.0-0.5 with 977 genes, 0.5-1.0 with 310 genes, 1.0-1.5 with 124 genes, 1.5-2 with 115 genes, more than 2 with 249 genes). Differences in vertebrate orthologous number between genes of these five groups were assessed. Figure 3 shows the mean values of orthologous numbers for each group of singleton genes in the fruit fly are dependent on the five ASEs level groups. Linear regression analyses revealed a positive relation between number of ancestral alternative splicing and the number of orthologous numbers in vertebrate genomes ($R=0.615\pm0.132$, $P<0.01$). Similar patterns were found among multicopy genes ($R=0.615$; $p<0.01$; Figure S3) and all genes ($R=0.707$; $P<0.01$; Figure S4) with the exception of the zebrafish. These results further reveal the evidence of a positive correlation between ancestral ASEs level and orthologous numbers in vertebrates.


Number of orthologs in Vertebrates
4.0
3.0
2.0
1.0
0.0
**
**
*
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in D.melanogaster

A


Number of orthologs in G. gallus
4.0
3.0
2.0
1.0
0.0
**
**
**
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in D.melanogaster

B



C

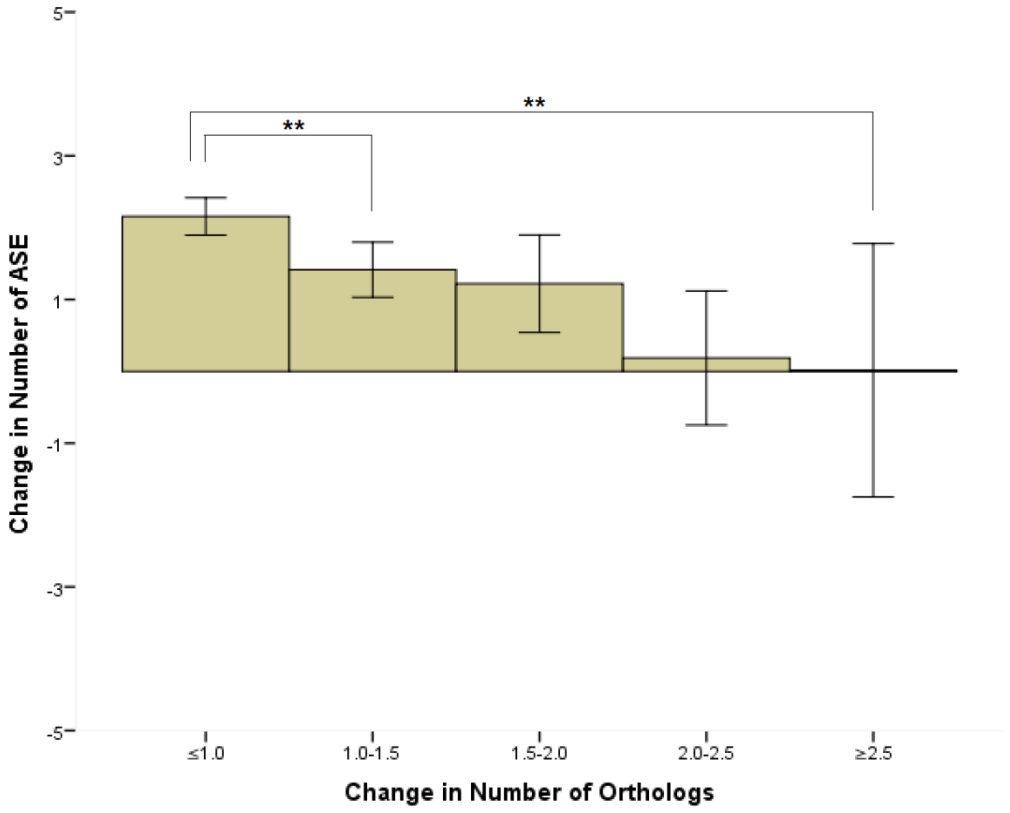
Figure 3 Comparison between different ASEs level in *D. melanogaster* and number of orthologous in vertebrates. Comparison of average number of ASE value of single copy genes in *D. melanogaster*, (A) average of vertebrates without *D. rerio* (Fish), (B) *G. gallus* (Chicken), (C) *H. sapien* (Human). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.

Higher orthologous numbers in vertebrates is linked to a slow-down in growth of alternative splicing levels

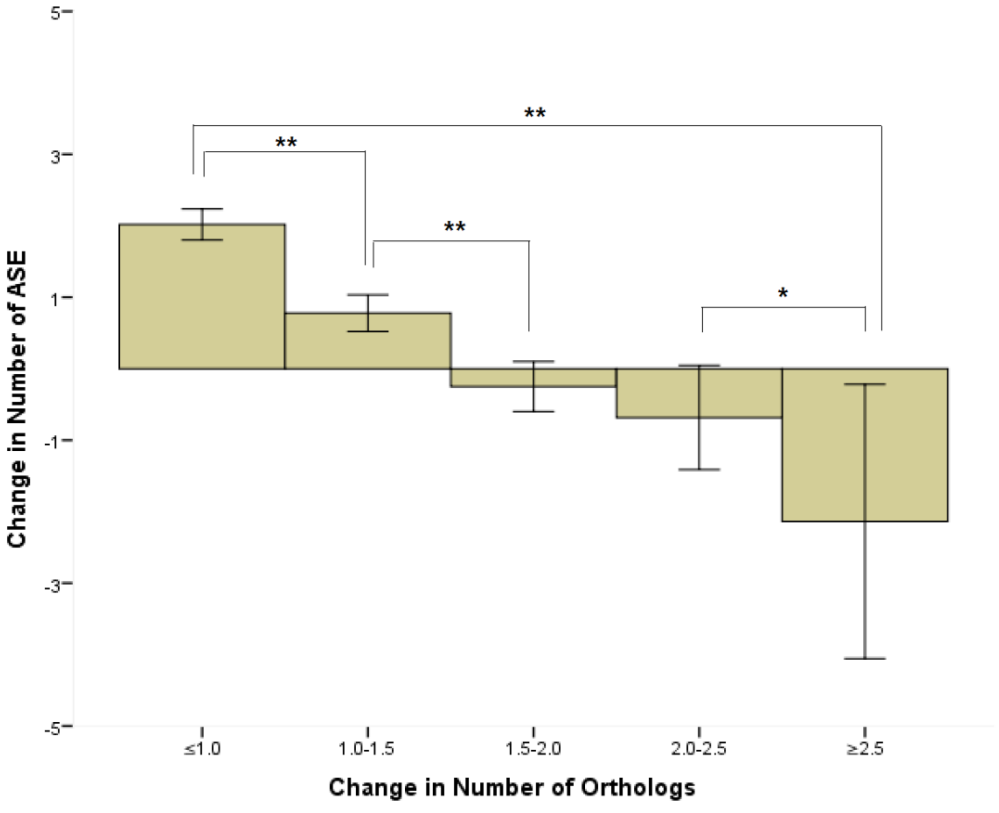
We next assessed if the positive relationship between the number of alternative splicing events in the invertebrate genome and the number of orthologous in vertebrate genomes is associated with a subsequent decrease in the rate of gain of new alternative splicing events as would be expected under a subfunctionalisation model.

For this, the data (singletons in fly) was divided into 5 groups by the orthologous numbers ratio: 0.0 (still singletons in vertebrates), 1.0 (two copies in vertebrates), 1.5-2.0, 2.0-2.5, >2.5 (three or more copies in vertebrates) and then we tested the correlation between the change in alternative splicing levels in the invertebrate to the vertebrate per gene (calculated as log 2 value of ASEs vertebrates per gene/ASEs invertebrates per gene) and the change in gene family size (assessed as the log 2 value of orthologous numbers in vertebrates/ orthologous numbers invertebrates).

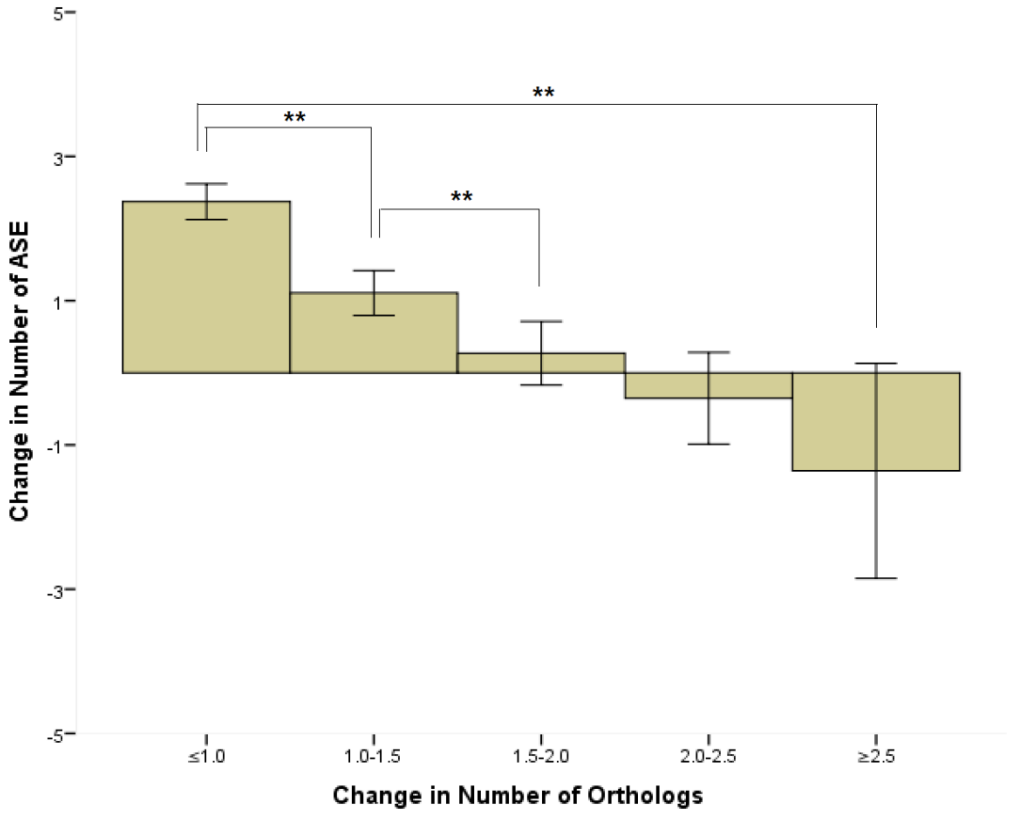
We found significant decreases in return rate of alternative splicing events from the vertebrate to the vertebrate genomes for genes with higher increases in orthologous numbers. This pattern was found in all species of vertebrates tested (average p= 0.002, chicken p=0.0002, human p=7.16E-07; Figure4). Focusing on the comparison of '0.0' group and the adjacent group '1.0', the significance of decreases (average p= 0.009, chicken p=3.89E-09, human p=1.99E-07) directly proved the degeneration of ASE in duplicates. Other decreases were shown between group '1.0' and group '1.5-2.0' in chicken (p=0.0006) and human (p=2.73E-05), and between group '2.0-2.5' and group '>2.5' in chicken (p=0.014). Calculation on genes with multiple copies and the collective results of both multiple copies and single copy show universal and consistent trend of reduction which supports the prediction of the sub-functionalization hypothesis. Most of the pairs of adjacent groups show the significant differences (FigureS11 and S12). These results show that higher increases in orthologous numbers in vertebrate genomes compared to invertebrates is associated with a decreased gain of alternative splicing events.


5
3
1
-1
-3
-5
**
**
Change in Number of ASE
≤1.0
1.0-1.5
1.5-2.0
2.0-2.5
≥2.5
Change in Number of Orthologs

A


5
3
1
-1
-3
-5
**
**
**
*
Change in Number of ASE
≤1.0
1.0-1.5
1.5-2.0
2.0-2.5
≥2.5
Change in Number of Orthologs

B

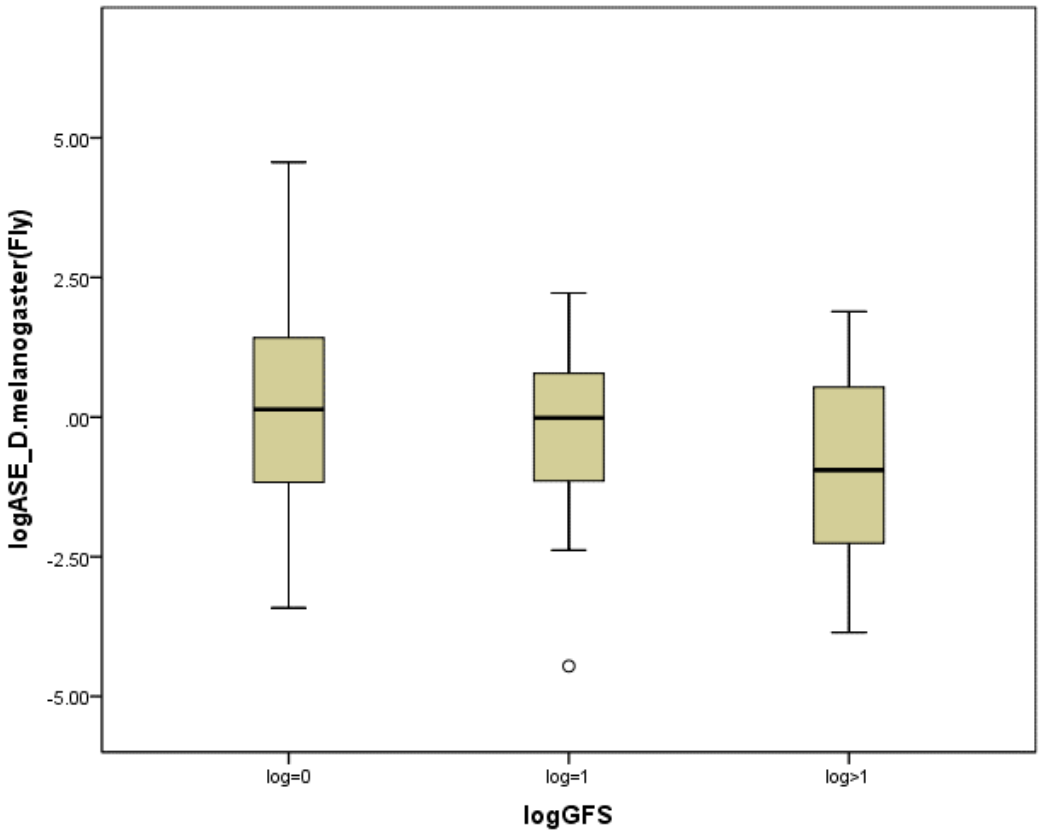


C

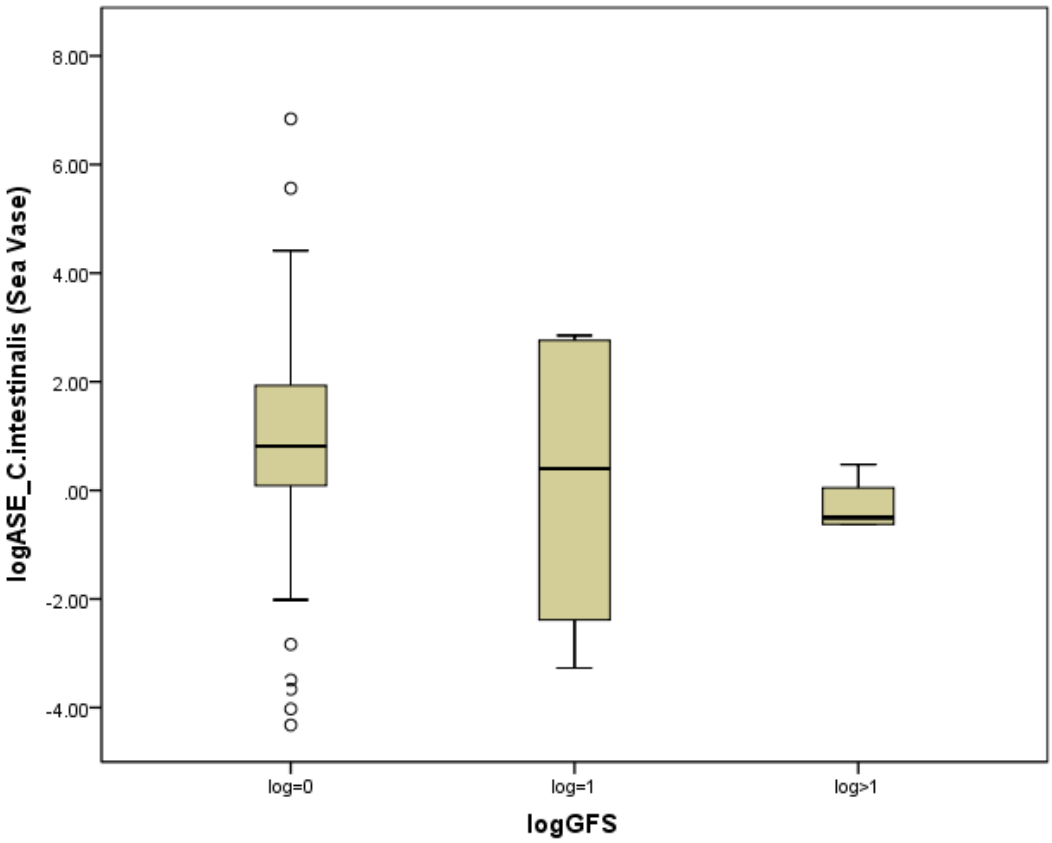
Figure 4 Change in Number of orthologous and Number of ASE between *D. melanogaster* and vertebrates. single copy genes in fly (A) average of vertebrates without *D. rerio* (Fish), (B) *G. gallus* (Chicken), (C) *H. sapien* (Human). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.

Higher orthologous numbers inside the invertebrates and vertebrates group have a similar slowdown trend of splicing increase

We compare the data inside the invertebrates (*C.elegans* singletons as an ancestral species) and vertebrates (*D. rerio* singletons as an ancestral species) groups to avoid the bias of our ancestor selection. Similarly, we tested the correlation between the log change in alternative splicing levels in the invertebrate and the vertebrate groups per gene and the log change in gene family size.


logASE_D.melanogaster(Fly)
5.00
2.50
.00
-2.50
-5.00
log=0
log=1
log>1
logGFS

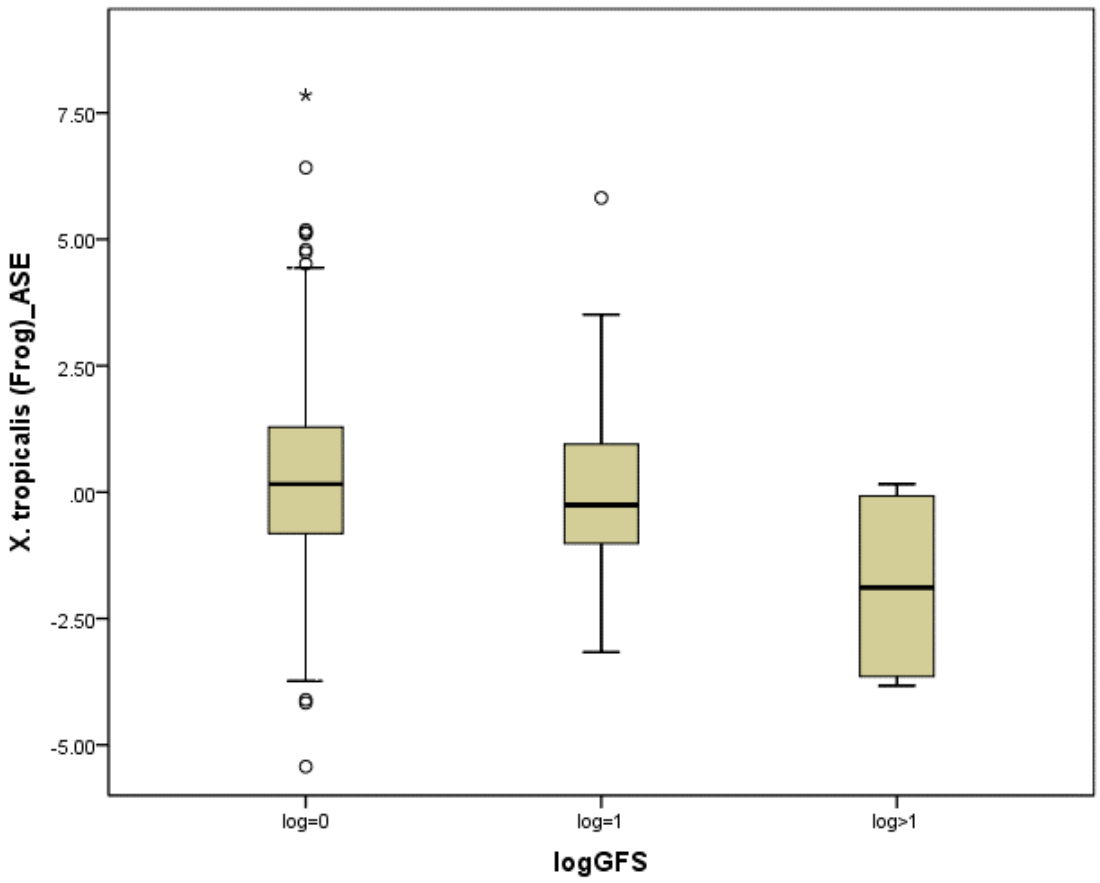
A.


logASE_C.intestinalis (Sea Vase)
8.00
6.00
4.00
2.00
.00
-2.00
-4.00
log=0
log=1
log>1
logGFS

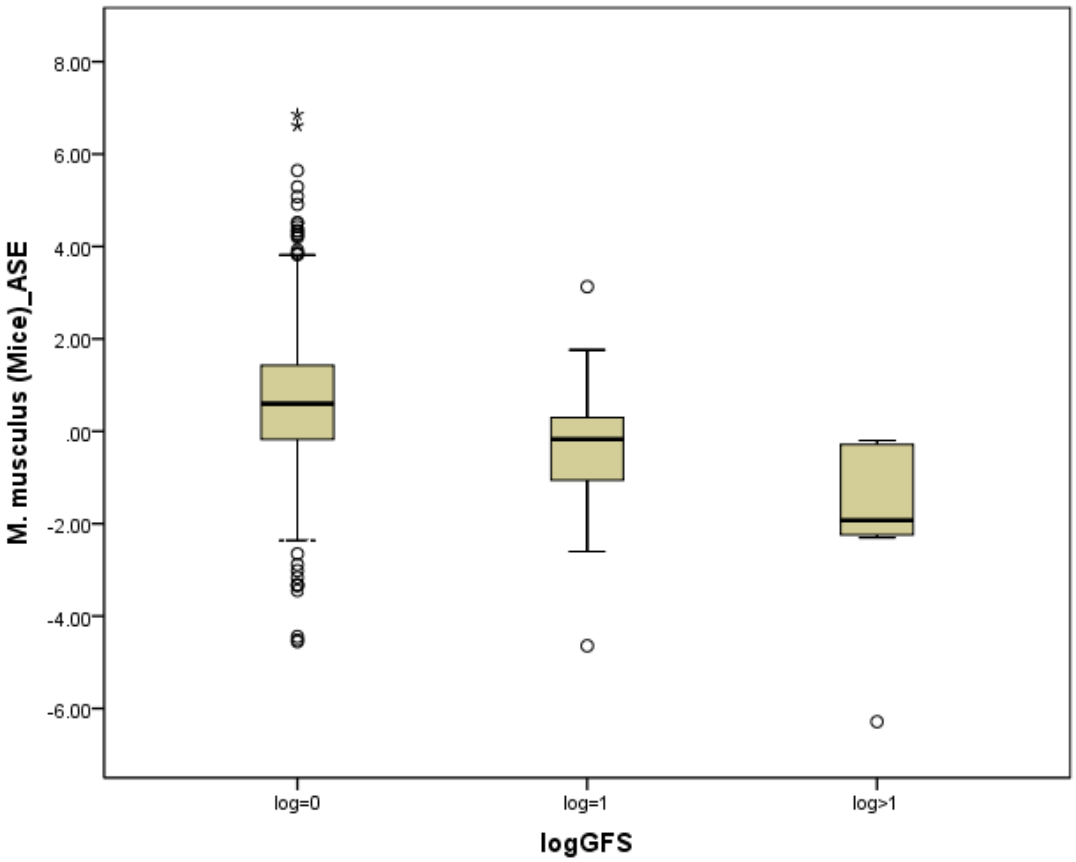
B.

Figure 5 Change in Number of orthologous and Number of ASE between *C.elegan* and other two invertebrates. single copy genes in *C.elegans* compare with (A) *D. melanogaster*, (B) *C. intestinalls*.

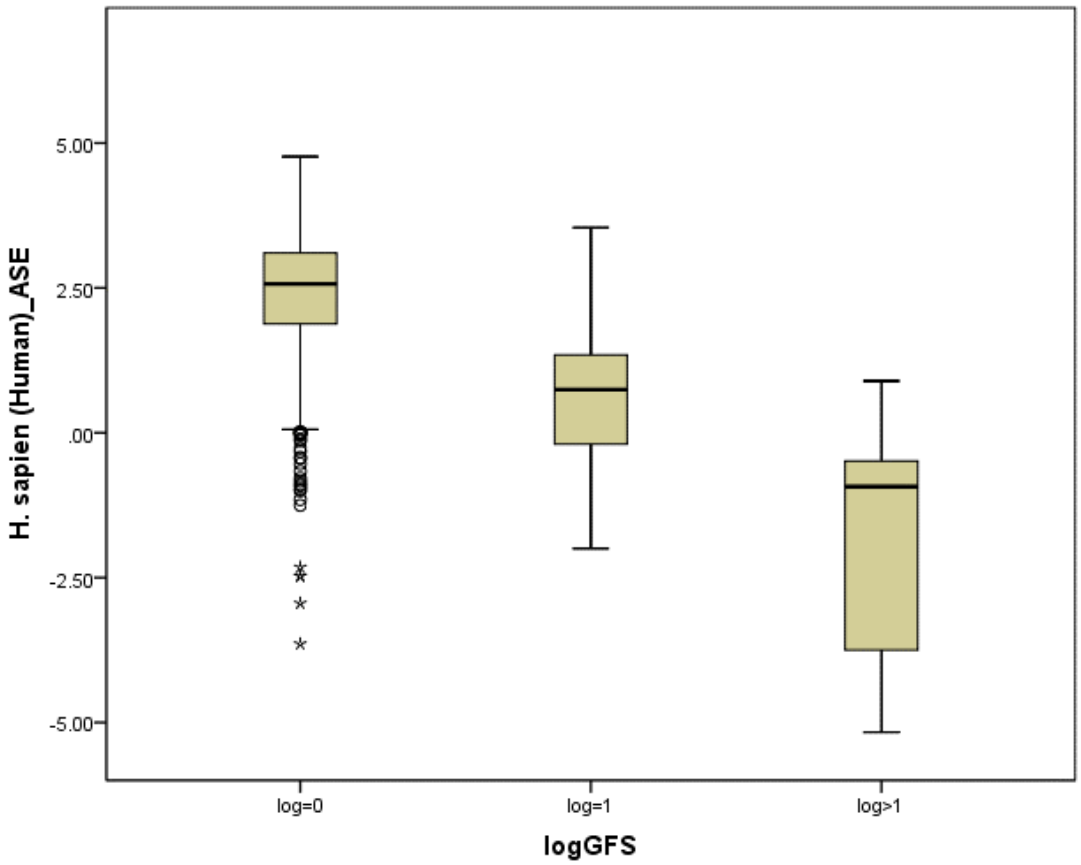
The slowdown of ASE was observed, which is similar with the results of the analogy between invertebrates and vertebrates. The data of fruit fly (Rate= -0.59, R2=0.98, p=0.006) is reliable and ciona vase (Rate= -0.41, R2=0.02, p=0.08) is a false results to prove the hypothesis. The results are might leaded by the shrink of our data size (2063 singletons in *C. elegan*, 94 and 122 valid orthologous data in *C. intestinal* and *D. melanogaster* respectively).



A.


8.00
6.00
4.00
2.00
.00
-2.00
-4.00
-6.00
M. musculus (Mice)_ASE
log=0
log=1
log>1
logGFS

B.


5.00
2.50
.00
-2.50
-5.00
H. sapien (Human)_ASE
log=0
log=1
log>1
logGFS

C.

Figure 6 Change in Number of orthologous and Number of ASE between *D. rerio* and other vertebrates. Single copy genes in *D. rerio* compare with (A) *X. tropicalis*, (B) *M. musculus*, (C) *H. sapiens*.

According to Figure1, *D. rerio* was chosen as the initial species to detect the change of other vertebrates GFS and ASE. . Lizard only has 22 valid data and the 493 singletons of fish and 489 singletons in the chicken species remain (Figure S16).Lizard, fish and chicken have 22, 493 and 489 valid data respectively, all insufficient for analytical purposes, therefore they are not presented in the main body of our essay. The results from frog, mice and human all have a very significant slowdown ($p<0.01$), but the R2 did not show the reliability of the outcomes. Thus this test of invertebrates and vertebrates enhances our suggestion of the relationship between GFS and ASE.

Another orthologous database to double confirm our conclusion

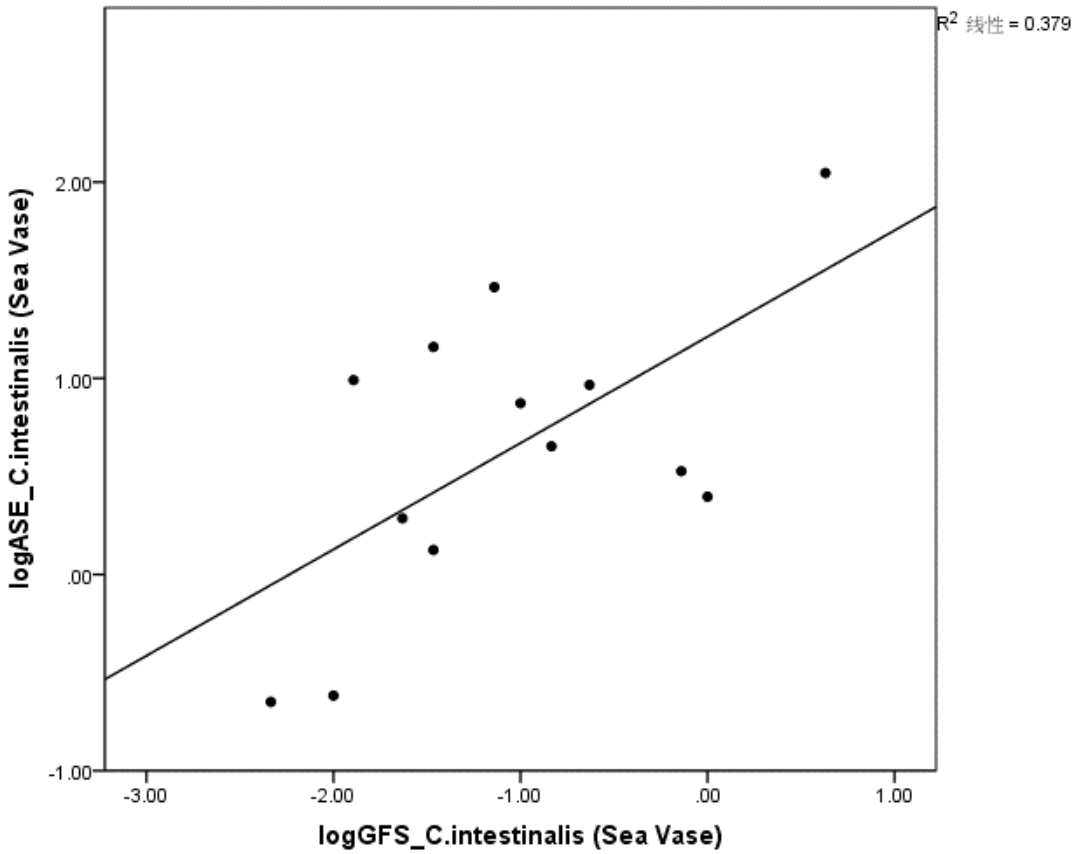
Considering the accuracy of our orthologous resource, data was downloaded from the Enzembl Biomart Orthologous instead of identifying orthologous by the Enzembl protein ID. Totally there are 5.29 million data of 10 species by adding *Ciona savignyi* into our research.

The reason we added *C. savignyi* is trying to increase the validity. *C. savignyi* is also one of the Ciona Genus and is actually a transitional species between invertebrates and vertebrates[33,34].

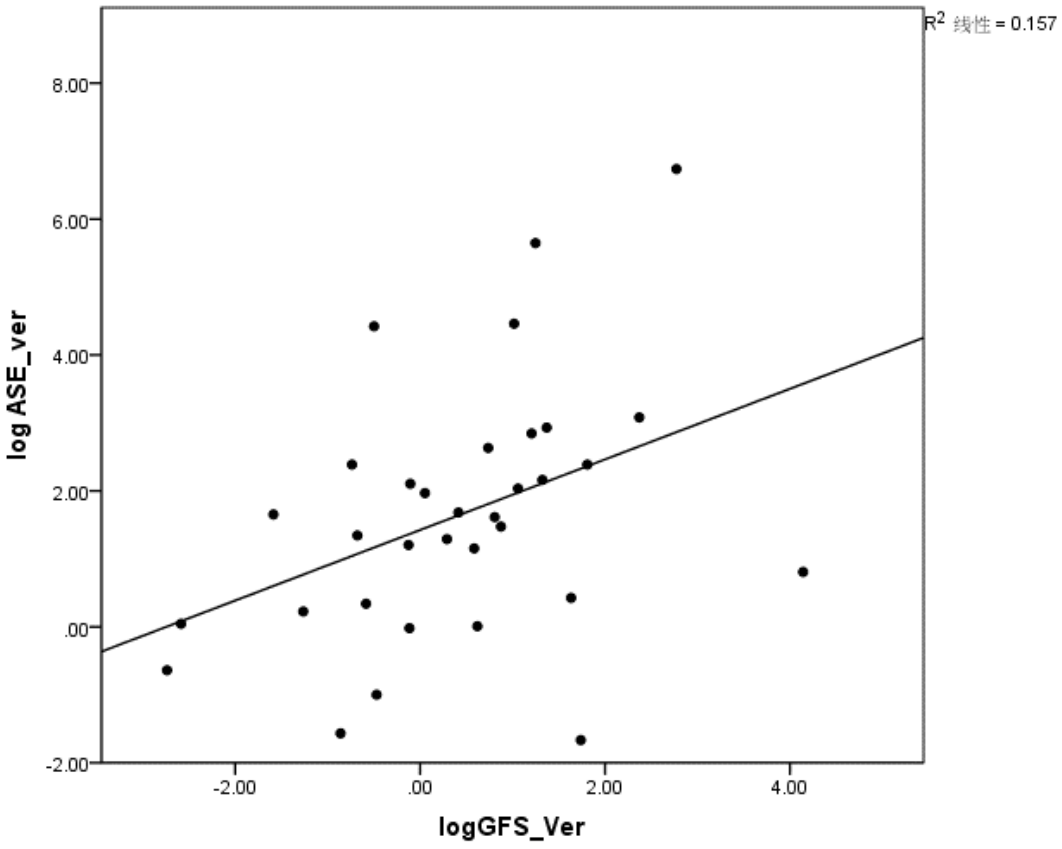
The log statistics of GFS and ASE changes based on the *C. elegan* was shown as following figures (Figure 7 and Figure S17). One essential point must be mentioned here is after all primary analysis (remove duplicates, filtered by our existing ASE database, log calculation), very few valid data is left for analysis from the 9 species (Table 3).

Table 3 **Experiment Species Valid Data List**

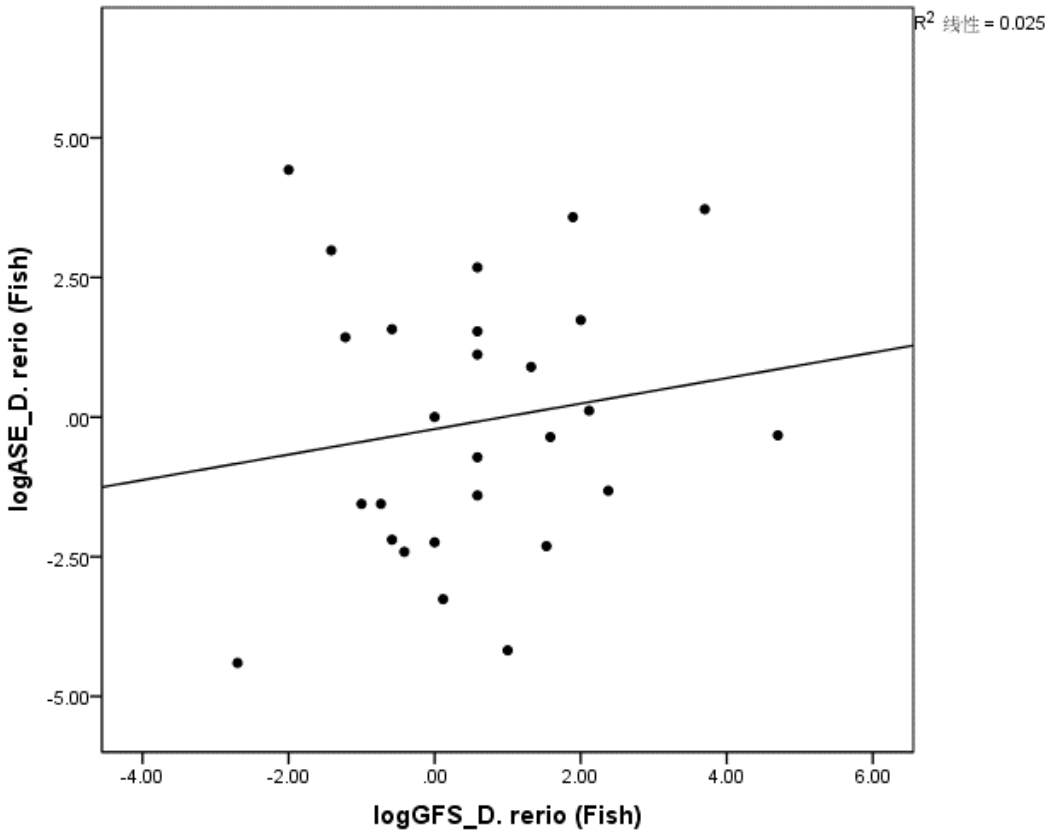
Species Name	Fly	Ciona1	Ciona2	Average	Fish	Frog	Lizard	Chicken	Mice	Human
Valid data	10	13	2	33	27	24	4	12	20	15


R² 线性 = 0.379
logASE_C.intestinalis (Sea Vase)
2.00
1.00
.00
-1.00
-3.00
-2.00
-1.00
.00
1.00
logGFS_C.intestinalis (Sea Vase)

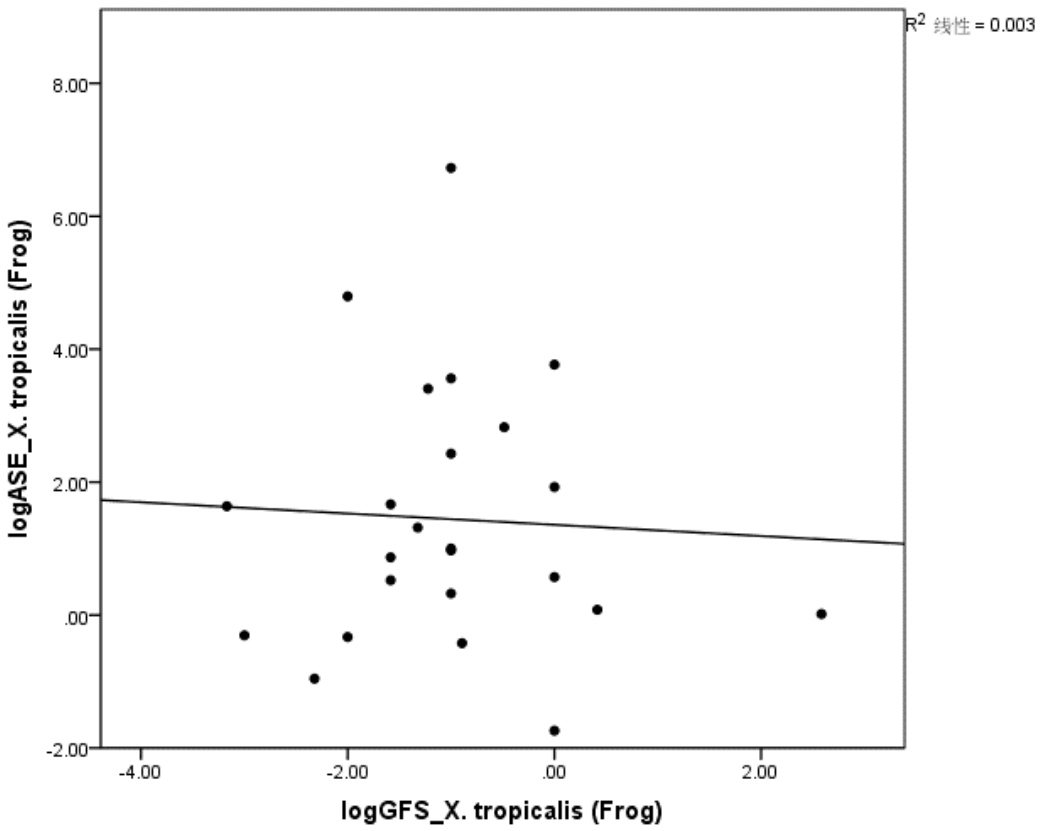
A.


R² 线性 = 0.157
log ASE_ver
8.00
6.00
4.00
2.00
.00
-2.00
-2.00
.00
2.00
4.00
logGFS_Ver

B.


R² 线性 = 0.025
logASE_D. rerio (Fish)
5.00
2.50
.00
-2.50
-5.00
-4.00
-2.00
.00
2.00
4.00
6.00
logGFS_D. rerio (Fish)

C.


R² 线性 = 0.003
logASE_X. tropicalis (Frog)
8.00
6.00
4.00
2.00
.00
-2.00
-4.00
-2.00
.00
2.00
logGFS_X. tropicalis (Frog)

D.

Figure 7 Change in Number of orthologous and Number of ASE between *C.elegan* , invertebrates and vertebrates. Single copy genes in *C.elegan* compare with (A) *C. intestinalis*, (B) average of Vertebrates, (C) *D. rerio*,(D) *X. tropicalis*.

We choose the species which have the most data to present the results. No fixed pattern was observed and the average of all vertebrates' data reveals a positive relationship rather than the expected slowdown. In the future, we planned to have a bigger database to avoid the bias.

Cell component genes have slightly smaller gene family size

Our project aimed to retest the relation between GD and gene functions domain. Furthermore, we want to observe the change of GD and AS in 9 species genome to uncover the role of gene functions during the evolution. In our project, we used the Gene Ontology Project data obtained from the Ensembl.org as a representation of the functions of the genes. The reason we choose GO Project as our dataset is due to the exhaustive functional resources it provides. The project cited the finding from 0.14 million papers and journals to build the dataset [35,36]. It combined the function-focused Gene Ontology (GO) and GO annotations together to provide the structured and classified dataset of over 6 million annotations. GO relies on the gene functions and gene transcripts to reveal the structure of the biological functions and their relationships.

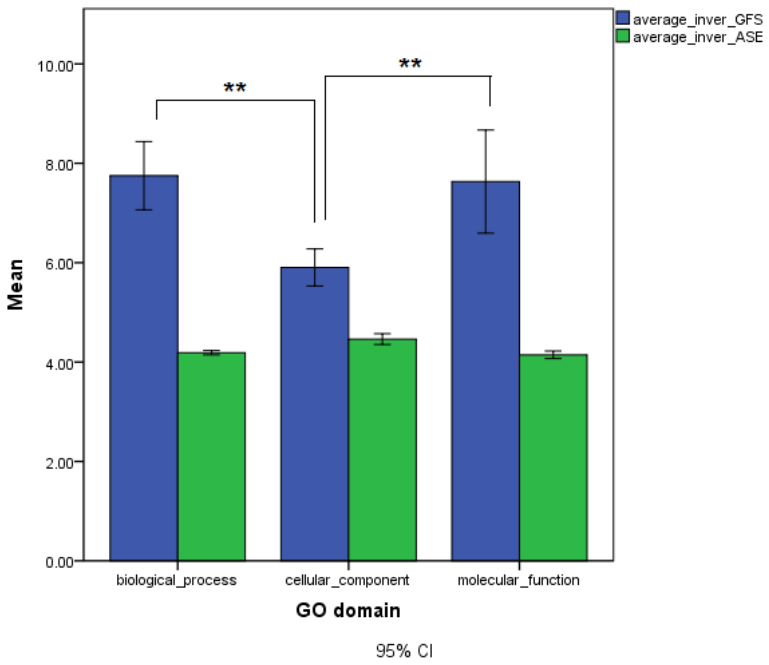
In our research design, GO ID was used as the marks of the "homology" instead of Ensembl protein ID in above study. There are three main GO domains: biological process, cellular component and molecular function. We compare the average GFS and ASE level of every GO domain to identify the general influence of functions' diversity and the change log data of them from invertebrates to vertebrates.

Firstly, we vertically compare the level of GFS and ASE in each species and the average of all invertebrates and vertebrates species (Figure 8 and S18). The GFS of the cellular-component genes

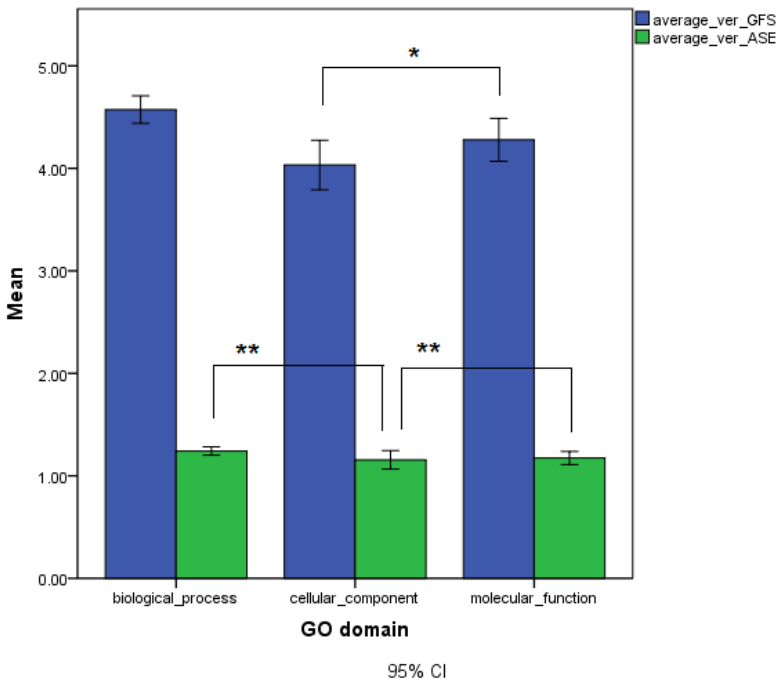
are always the lowest among the 3 types and nearly half show the significant difference. In the contrast, the ASE of cellular component was the highest in few species, which means the production of different transcripts in component process is relatively more dependent on alternative splicing.

Table 3 **Data amount of three GO domain**

Go domain	Total	Biological process	Cellular component	Molecular function
Data	5913	3302 (56%)	807 (14%)	1804 (30%)



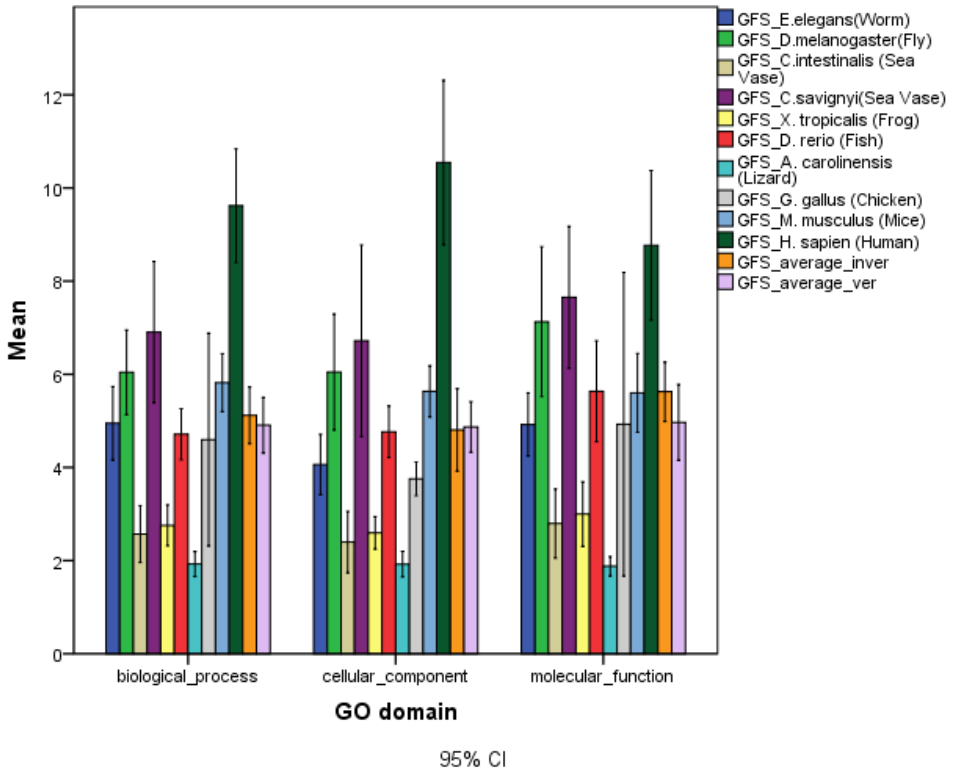
A.



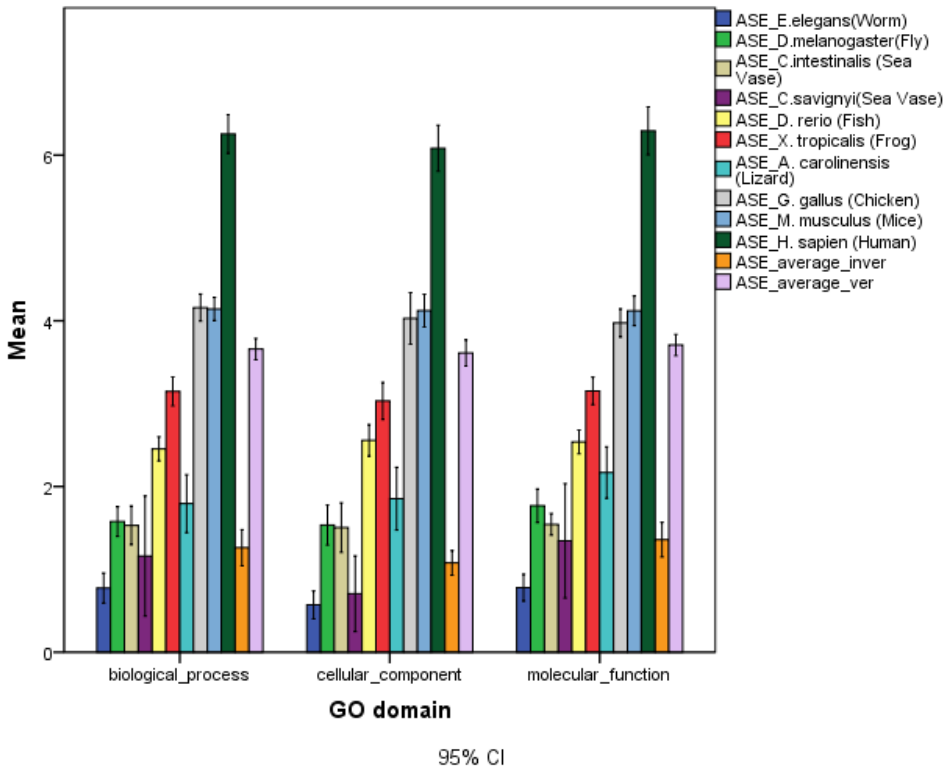
B.

Figure 8 Compare between the GFS and ASE level of three genetic domains (biological process cellular component and molecular function) (A) average of invertebrates, (B) average of vertebrates. *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.

The followed step is horizontal analysis between 10 species (Figure 9). Although the level of GFS increased from frog to human, overall GFS did not show a fixed pattern of correlation with the complexity of species. Especially the average of invertebrates is even higher than the vertebrates mean value. In contrast, the ASE level has a significant increase which basically followed the evolution tree we built at the beginning of the research and prove the Nuno's results[28].



A.



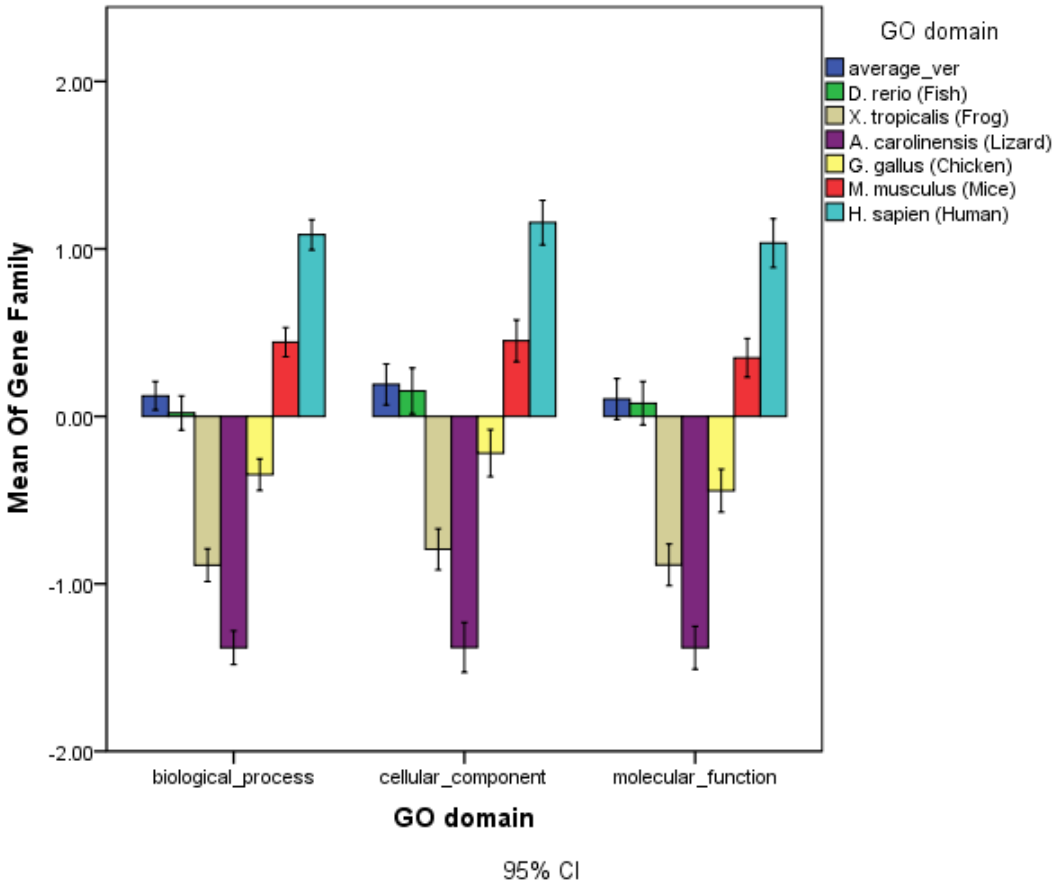
B.

Figure 9 Overall review of the GFS and ASE level of three genetic domains (biological process cellular component and molecular function) in different species (A) GFS level in 10 species, (B)

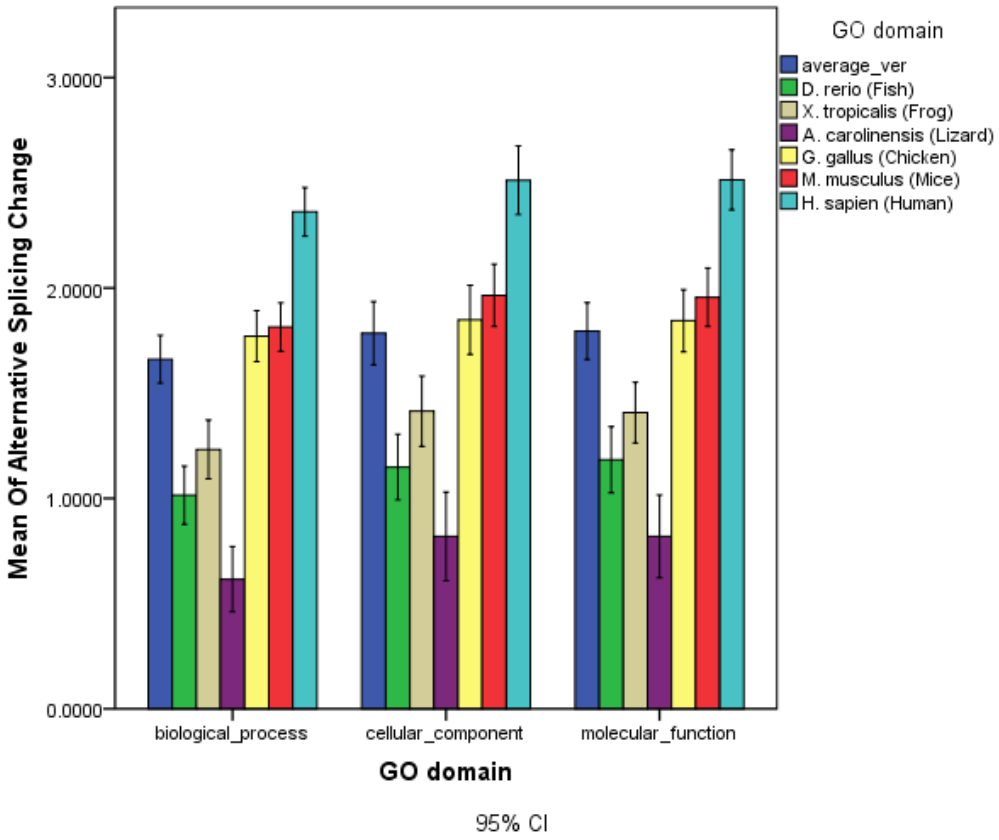
ASE level in 10 species.

Gene family size and splicing has a faster increasing rate in the advanced vertebrates

Thirdly, the log of change of GFS and ASE from invertebrates to vertebrates was calculated to detect the function effect during the evolution. In general Figure 10 prove the results above that the change of ASE has a higher value in the complex species in all three GO domains ($p<0.01$). And the GFS figure shows that all species have a significantly ($p<0.01$) positive relations with the GFS change rate. However, the analysis of zebrafish and lizard do not present significant pattern, potentially due to insufficient valid data.



A.

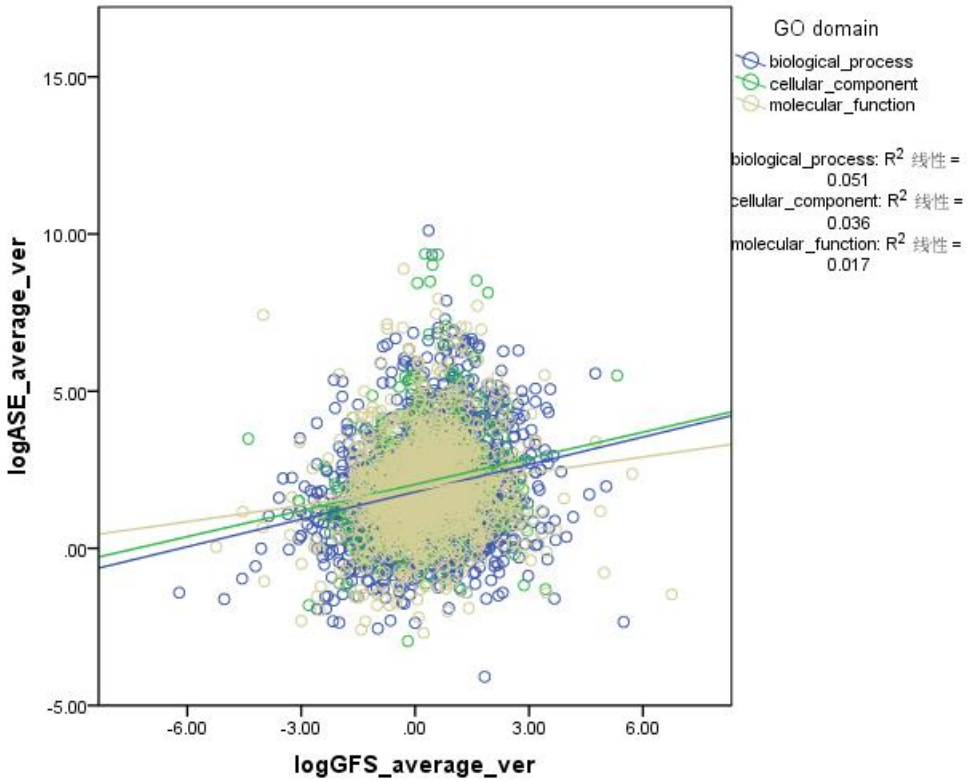


B.

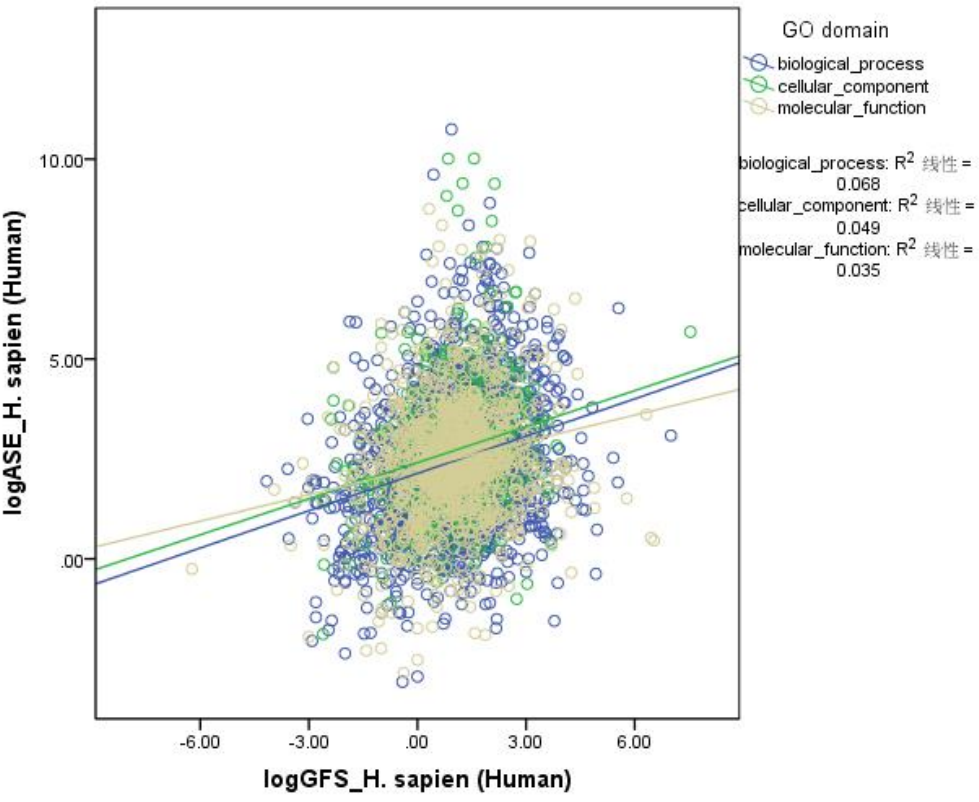
Figure 10 Overall review of the log value of GFS and ASE change level from invertebrates of three genetic domains (biological process cellular component and molecular function) in different species (A) GFS change level in 10 species, (B) ASE change level in 10 species.

Functional differentiation has a minor effect on the change of gene family size and alternative splicing

Finally, we compare the co-relation between ASE change rate and GFS change rate to prove the conclusion on front. Generally, the ASE has a slightly positive association with the increase of GFS (Figure 11 and S20, $p < 0.05$). Go deep into every single vertebrate species GFS and ASE relation, the increasing trend also emerged in our results. And it can also be observed that the cellular component is the fastest ASE growth and molecular function is the lowest ASE growth. The results might be affected by the noise of multiple genes in ancestors.



A.

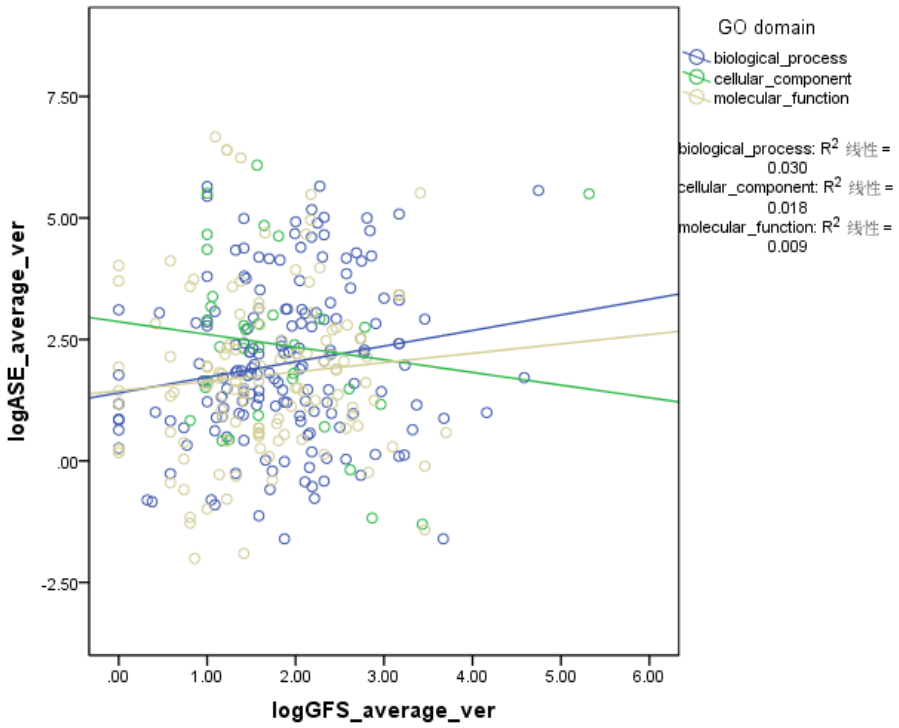


B.

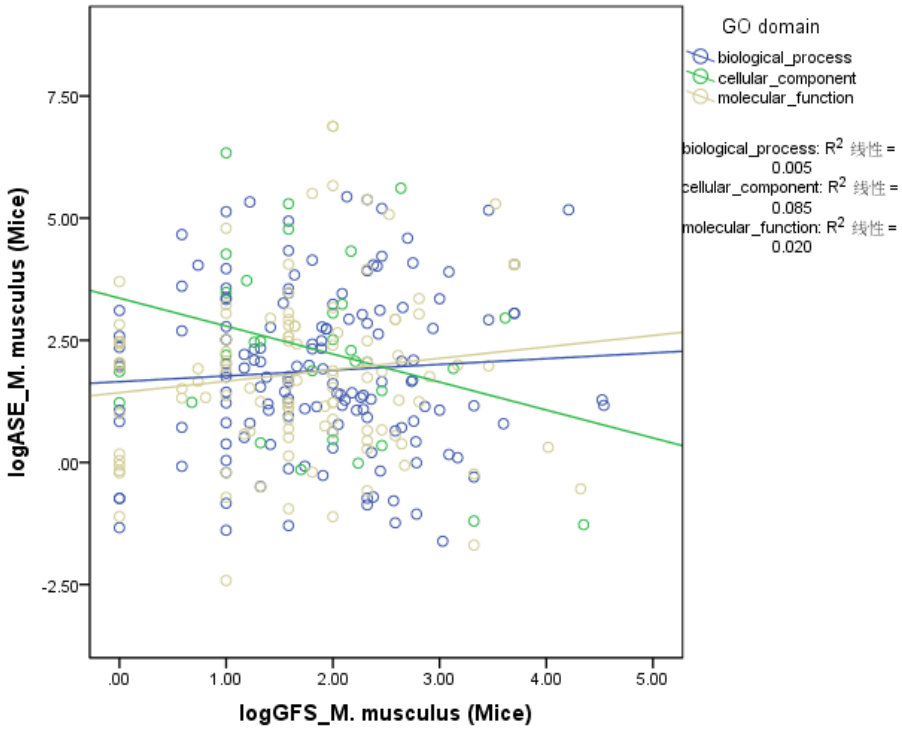
Figure 11 Co-relation between the log value of GFS and ASE change level from invertebrates of three genetic domains (biological process cellular component and molecular function) in

different species (A) Mean value of all vertebrate species, (B) *H.sapien*.

Therefore our study concentrates the change of singleton in invertebrates (Figure12 and S21). The cellular component gene of mice, frog, and chicken, reveal the reduce trend of ASE ($p<0.05$). While most of the process and molecular function gene still has the positive co-relation. This might provide a pattern that the cellular component gene are more conservative to develop more neo-function, in which situation that gene duplications would be used to replace the function of ancestral ASEs.



A.



B.

Figure 12 Co-relation between the log value of GFS and ASE change level from invertebrates singletons of three genetic domains (biological process cellular component and molecular function) in different species (A) Mean value of all vertebrate species, (B) *M.musculus*.

DISCUSSION

Our analysis aimed to test the hypothesis that the effect of pre-duplication ancestral alternative splicing on the choice of fates of the duplicated genes. First, we measured the ASEs and orthologous family size based on the data gathered from Ensembl dataset (see in methods). RNA-seq (has the ability to produce millions of short sequence read [37-39])and ESTs (randomly selected sequence reads derived from cDNA[40]) are both effective methods to obtain the splicing data. We choose the ESTs in our test because it yields longer transcript segments. And also ESTs are more comprehensive in terms of species than the RNA-seq data [41] . However, the proportion of valid sample in EST is less than our exception in some species, exampled in the data gathered from lizard. RNA-seq is planned to be applied subsequently to update our ASEs dataset. In term of identification of orthologous families' sizes, protein ID was regarded as the annotation. Few protein families, such as ENSFM00740001589073 in every species, have outliner values. The reason of this phenomenon is still unknown nowadays, it might be a bias induced by our identification methods and should be fixed in our future research.

Secondly, we found that the duplication with ancestral ASEs has a significantly larger mean of family size. The further detailed test reveals the positive relationship of this two: more ASEs in the ancient would results bigger orthologous family size. The results focus on building a dynamic 'pattern' between 'correlation' (ancients with ASEs or not) and 'results' (retention of duplicates or not). This pattern is commonly noticed on the level of whole ancestral genomes. We can suggest that during the selection of duplication fates under the environmental stress, by the requirements of some functions, bias was induced by AS[42-45]. We noticed that the affecting factors of pre-duplications have not been considered: the larger ancestral gene family sizes might lead to larger vertebrate gene family sizes. The results provide evidence for the retention of the duplicates leading to the loss of the ASEs, which is consistent with sub-functionalization model; rejecting the hypothesis that ASEs function merely as a marker of selection for larger gene family sizes.

Then, the analysis between the invertebrates and vertebrates confirms our on-front finding. While when we tried to eliminate bias caused by the orthologue calculation through a new resource, the size of the dataset is not enough to build a reliable conclusion.

Finally, the functional analysis did provide us the matter of species complexity rather than any significant difference between different GO domains in the same species. It confirms the conclusion of Nuno's study [28]. In our further research, we planned to divide the domains into detailed groups to recognize the effect of the functions. The GO slim database is a suitable database to do the function clarification [35,36]. GO slim is the subsets of GO database. It is the summary of the GO term description and makes the comparison easier.

Overall, findings presented here show that ancestral alternative splicing status inferred from invertebrate genes is associated with higher numbers of orthologous numbers in vertebrate genomes. More pronounced increases in orthologous numbers are in turn associated with a decrease in the gain of alternative splicing events observed in the vertebrate genes compared to their invertebrate orthologous. Thus, we conclude that our results are in agreement with expectations under a model whereby alternative splicing increases the likelihood of duplicate gene retention after whole genome duplication events.

In the future, increasing ASE database size would be the prior mission to avoid any bias in our study. Then base on Lu's research[13], the other 5 types of ASE would be calculated (random 3s, 5s, 3s5s, Exon skipping or cassette exon, Intron Retention[46]) and compared with the level of GFS. Meanwhile, the exons types would be considered[47]. Will the alternative exons contribute more than constitutive exons during the retention of gene duplications? Thirdly, as we mentioned above, GO slim data will be used to update our functional research. Other factors like age [48] of genes or some particular genes would also be tested.

METHODS

Identification of ASEs

Genome sequences and annotations was downloaded from Ensembl and full mRNA and EST sequences from UniGene[49]. To build a comparable AS database for 18 eukaryotic genomes, random samples of 10 transcripts from each gene's pool of mRNA and EST sequences (modified from refs. [50,51]). In order to compare the expression across species, we grouped library to 10 common organ level according to BodyMap-Xs[52]for 10 species and employed a random sampling to reconstruct the library, in which we randomly selected 10,000 ESTs for 100 times from the libraries of each organ, then times of each gene presented in this randomly ESTs library with one million ESTs were used as a proxy of gene expression. Data were extracted as different type of ASEs of each Gene IDs.

Ideyification of size of orthologous families

1.Gene IDs and Protein family IDs information were downloaded from Ensembl BioMart [53,54]. The size of paralog families was calculated from the frequency of the protein family IDs. The ASEs level of each paralog family was obtained from the sum-up of ASEs data of all the genes in this family. The ASEs data paralogous of each species was combined to identify the orthologous families and the ASEs data of them in each species.

2. Direct Gene Orthologues information of 10 species were download from Ensembl BioMart[54,55] to avoid the bias caused by method 1,

Linear regressions and T test

For the correlations between the size of Othologous family and ASEs level in invertebrates, a linear regression was applied to the dataset. And T test was used to the detailed comparison between the adjacent groups.

Tools

Excel and R were used for data statistic and R and SPSS were sued for the plotting.

REFERENCE

1 Innan, H. K., F. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* **11**, 97-108 (2010).

2 Lynch, M., Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459-473 (2000).

3 Ohno, S. Evolution by gene duplication. (1970).

4 Soskine, M. T., D. S. Mutational effects and the evolution of new protein functions. *Nature Reviews Genetics* **11**, 572-582 (2010).

5 Chow LT, G. R., Broker TR, Roberts RJ. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**, 1-8 (1977).

6 Berget SM, M. C., Sharp PA Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 3171-3175 (1977).

7 Alt FW, e. a. Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 30 ends. *Cell* **20**, 293-301 (1980).

8 Early P, e. a. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20**, 313-319 (1980).

9 Edger P, P. J. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Re* **17** (2009).

10 Konrad G., e. a. Fungal Alternative Splicing is Associated with Multicellular Complexity and Virulence: A Genome-Wide Multi-Species Study. *DNA Res* **21**, 27-39 (2014).

11 Pan, Q. e. a. Deep Surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413-1415 (2008).

12 Zhang C, e. a. Evolutionary character of alternative splicing in plants. *Bioinform Biol Insights* **9**, 47-52 (2015).

13 Lu, C. e. a. Comparative and Functional Analysis of Alternative Splicing in Eukaryotic Genomes. *PhD graduation essay* (2012).

14 Bush SJ, C. L., et al. Alternative splicing and the evolution of phenotypic novelty. *Philos Trans R Soc Lond B Biol Sci* **5** (2017).

15 Lu, C. e. a. Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity. *Mol. Biol. Evol* **31**, 1402–1413 (2014).

16 Eva Schad, P. T. a. H. H. The relationship between proteome size, structural disorder and organism complexity. *Genome Biology* **12** (2011).

17 Marshall, A. N. e. a. Alternative Splicing and Subfunctionalization Generates Functional Diversity in Fungal Proteomes. *Ed. Joseph Heitman. PLoS Genetics 9.3 Web* (2015).

18 S.T.Kariyazono, e. a. Genetic diversity of fluorescent protein genes generated by gene duplication and alternative splicing in reef-building corals. *Zoological Letters* **1** (2015).

19 Su.Z, e. a. Evolution of alternative splicing after gene duplication. *Genome Res* **16**, 182-189 (2006).

20 Naama M k., D. L. I. Y. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nature Genetics* **37**, 588-589 (2005).

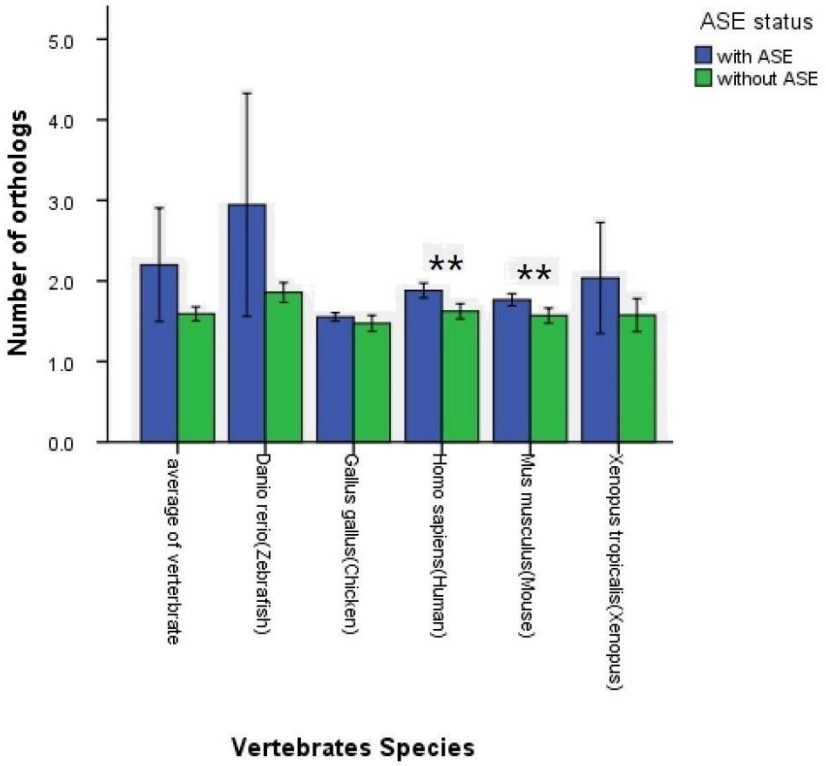
21 Koppelman NM, L. D., Yanai I. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat Genet* **37**, 588-589 (2005).

22 Roux, J., Rechavi, M.R. Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication. *Genome Res* **21**, 357-363 (2011).

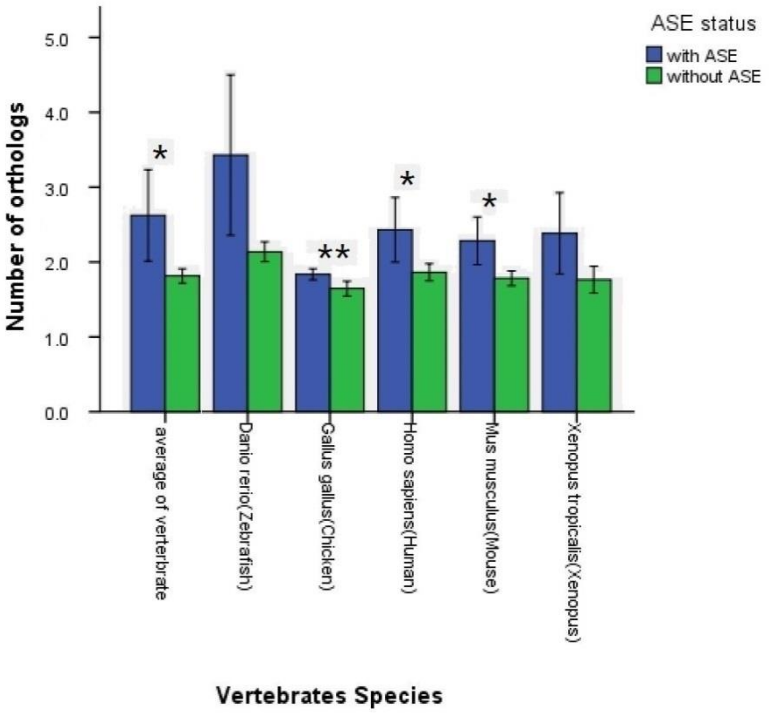
23 MacLean, D. W., Meedel, T. H. & Hastings, K. E. M. Tissue-specific alternative splicing of ascidian troponin I isoforms - Redesign of a protein isoform-generating mechanism during chordate evolution. *Journal of Biological Chemistry* **272**, 32115-32120 (1997).
24 Yu, W. P., Brenner, S. & Venkatesh, B. Duplication, degeneration and subfunctionalization of the nested synapsin-Timp genes in Fugu. *Trends in Genetics* **19**, 180-183, doi:10.1016/s0168-9525(03)00048-9 (2003).
25 Lister, J. A., Close, J. & Raible, D. W. Duplicate mitf genes in zebrafish: Complementary expression and conservation of melanogenic potential. *Developmental Biology* **237**, 333-344 (2001).
26 Altschmied, J. *et al.* Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish. *Genetics* **161**, 259-267 (2002).
27 Misook Ha, E.-D. K. a. Z. J. C. Duplicate genes increase expression diversity in closely related species and allopolyploids. *PNAS* **106**, 2295-2300 (2009).
28 Nuno L. Barbosa-Morais, e.-a. The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science* **338**, 1597-1593 (2012).
29 Chen L, e. a. Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts. *Hum Mol Genet* **20**, 4422-4429 (2011).
30 Lu Chen, e. a. Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity. *MBE* (2014).
31 Letunic I, B. P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **8**, 242-245 (2016).
32 Param Priya Singh , J. A., Hervé Isambert Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLOS Computational Biology* (2015).
33 Pearse, H. A. R. F. V. N. G. E. Localization of somatostatin-, substance P- and calcitonin-like immunoreactivity in the neural ganglion of Ciona intestinalis L. (Ascidiaceae). *Cell and Tissue Research* **202**, 263-274 (1979).
34 Leffler EM, B. K., et al. Revisiting an old riddle: what determines genetic diversity levels within species? *Plos Biology* **10**, doi:doi: 10.1371/journal.pbio.1001388 (2012).
35 Harris MA, C. J. e. a. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**, 258-261 (2004).
36 al, E. C. M. M. e. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res* **32**, 262-266 (2004).
37 Wang Z, e. a. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).
38 Robertson G, e. a. De novo assembly and analysis of RNA-seq data. *Nat Methods* **7**, 909-912 (2010).
39 Z, M. J. A. W. Next-generation transcriptome assembly. *Nature Reviews Genetics* **12**, 671-682 (2011).
40 Nagaraj N, e. a. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* **8**, 548 (2011).
41 Boguski, M. e. a. dbEST--database for expressed sequence tags *Nat Genet* **4**, 332-333 (1993).
42 Kondrashov FA, e. a. Selection in the evolution of gene duplications. *Genome Biology* **3**, Research (2002).
43 Otto SP, Y. P. The evolution of gene duplicates. *Ady Genet* **4**, 451-483 (2002).
44 Seoighe C, W. K. Yeast genome evolution in the post-genome era. *Curr Opin Microbiol* **2**, 548-554 (1999).
45 Guillemaud T, e. a. Quantitative variation and selection of esterase gene amplification in Culex pipiens. *Heredity* **83**, 87-99 (1999).
46 YAN WANG, J. L., et al. Mechanism of alternative splicing and its regulation. *Biomed Rep* **3**, 152-158 (2015).
47 Eyras, M. P. a. E. Differentiated evolutionary rates in alternative exons and the implications for splicing regulation. *BMC Evol Biol* **6**, doi:doi: 10.1186/1471-2148-6-50 (2006).

48 Hugo V. S. Rody, G. J. B., et al. Both mechanism and age of duplications contribute to biased gene retention patterns in plants. *BMC Genomics* **18**, doi:doi:10.1186/s12864-016-3423-6. (2017).
49 Sayers, E. W. e. a. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **37**, D5-D15 (2009).
50 Brett, D., Pospisil, H., Valcarcel, J., Reich, J. & Bork, P. Alternative splicing and genome complexity. *Nature Genetics* **30** (2002).
51 Kim, E., Magen, A. & Ast, G. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research* **35**, 125-131 (2007).
52 Ogasawara, O. e. a. BodyMap-Xs: anatomical breakdown of 17 million animal ESTs for cross-species comparison of gene expression. *Nucleic Acids Research* **34**, D628-D631 (2006).
53 Kasprzyk A, e. a. Ensmart: a generic system for fast and flexible access to biological data. *Genome Res* **14**, 160-169 (2004).
54 Haider, S. e. a. BioMart Central Portal-unified access to biological data. *Nucleic Acids Research* **37**, W23-W27 (2009).
55 HERRERO, J. How to get all the orthologous genes between two species. (2009).

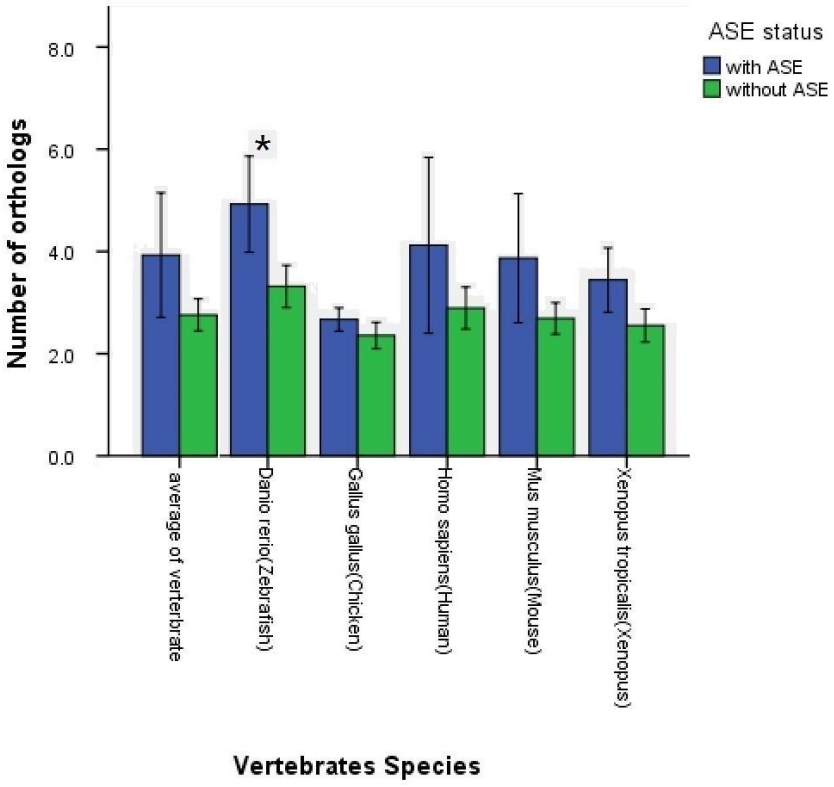
SUPPLYMENTS



A



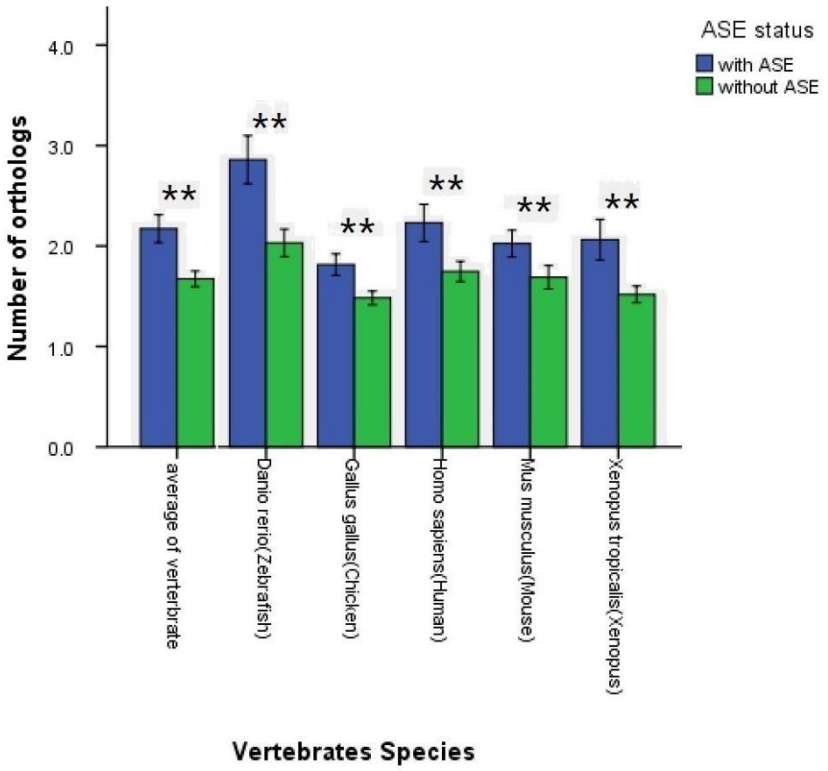
B



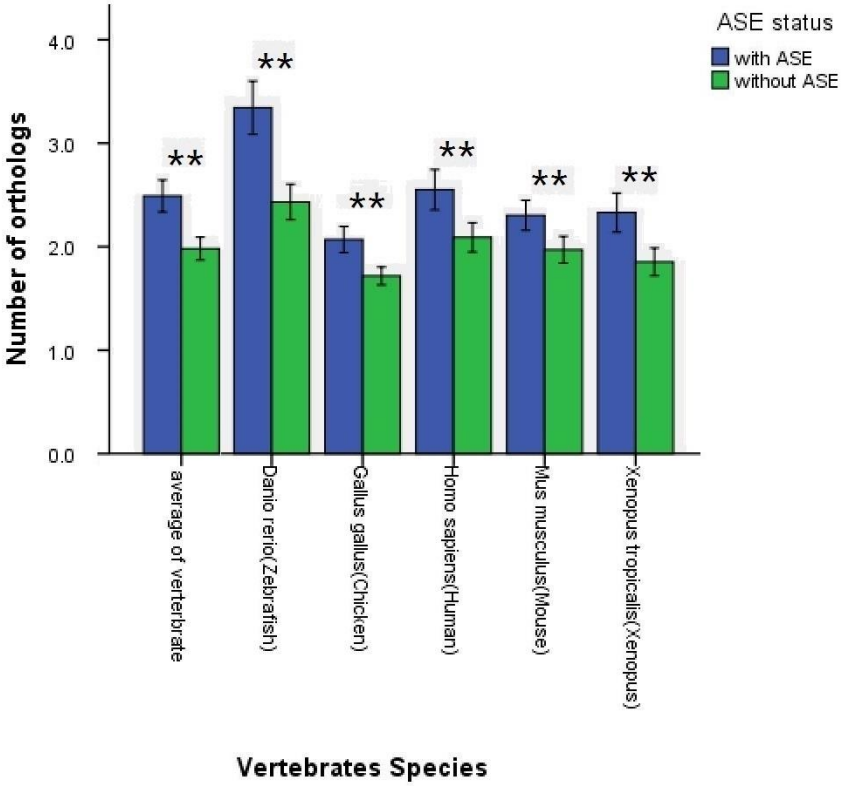
C

Figure S1 Comparison of number of orthologous between two gene categories (with and without ASE in *C. intestnalis*) in all vertebrates species (single copy (A) all genes (B) and

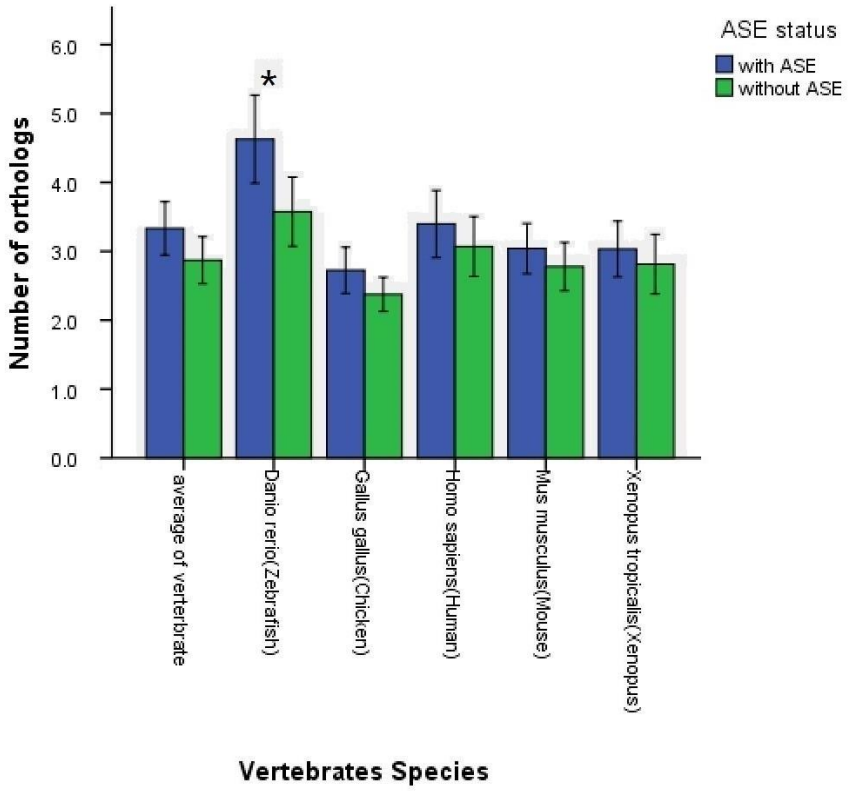
multiple copies (C). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.


ASE status
with ASE
without ASE
Number of orthologs
4.0
3.0
2.0
1.0
0.0
**
average of verterbrate
Danio rerio(Zebrafish)
Gallus gallus(Chicken)
Homo sapiens(Human)
Mus musculus(Mouse)
Xenopus tropicalis(Xenopus)
Vertebrates Species

A

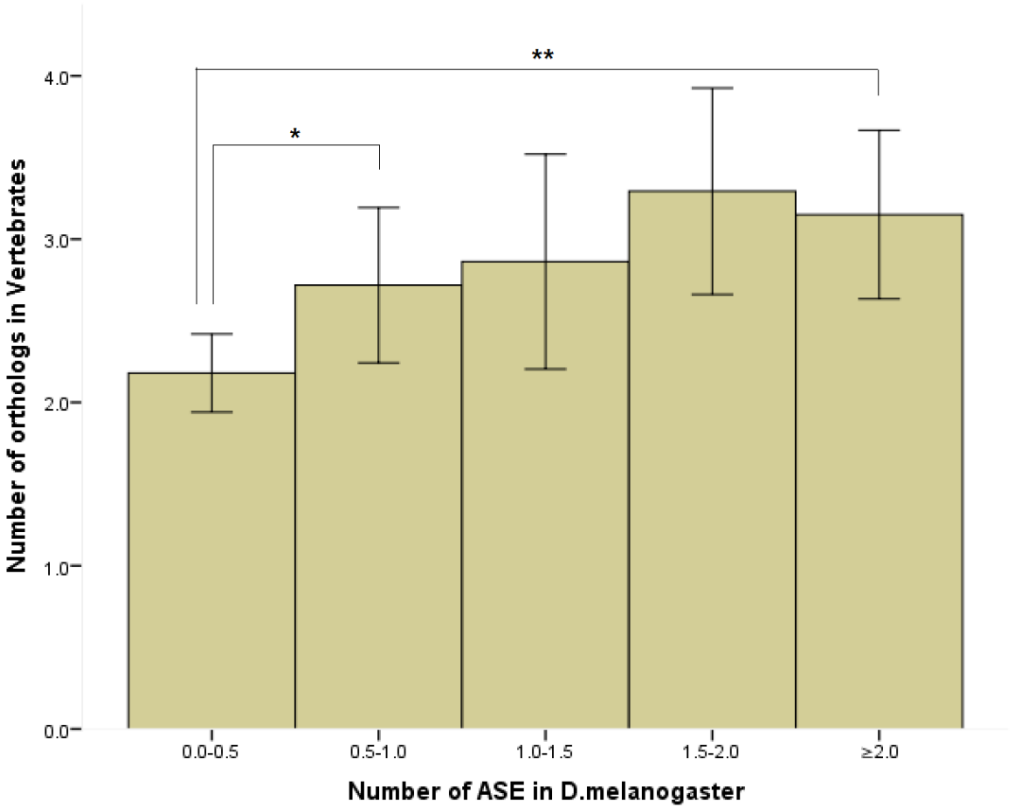

ASE status
with ASE
without ASE
Number of orthologs
4.0
3.0
2.0
1.0
0.0
**
average of verterbrate
Danio rerio(Zebrafish)
Gallus gallus(Chicken)
Homo sapiens(Human)
Mus musculus(Mouse)
Xenopus tropicalis(Xenopus)
Vertebrates Species

B

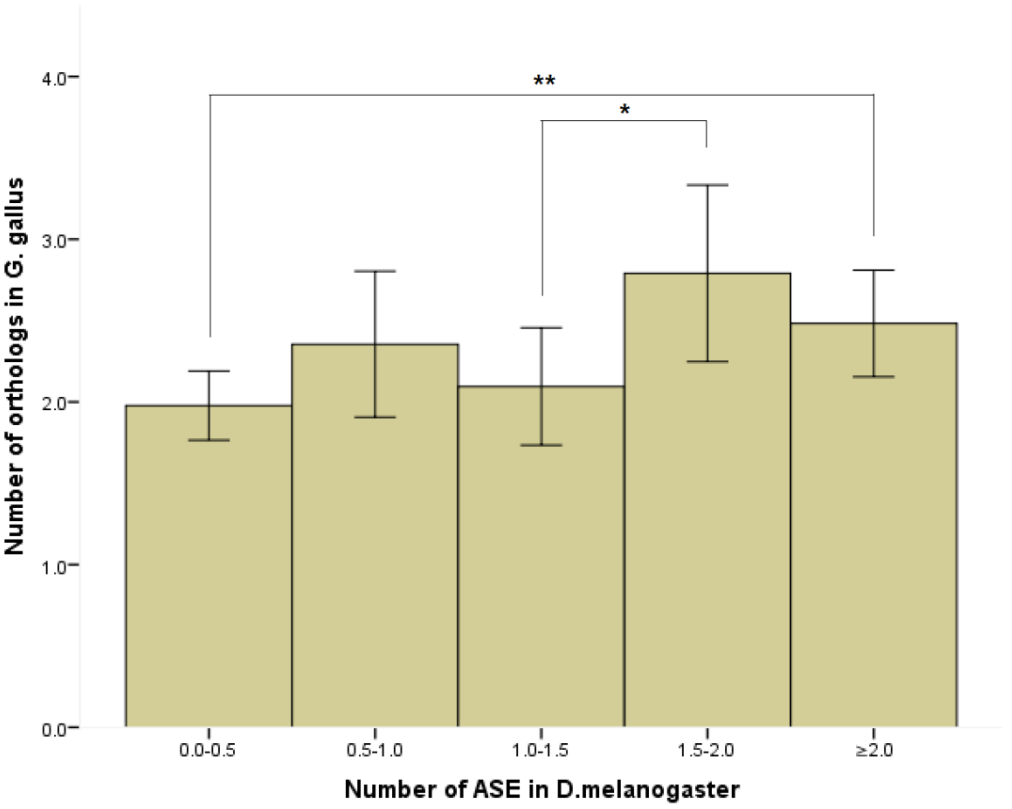


C

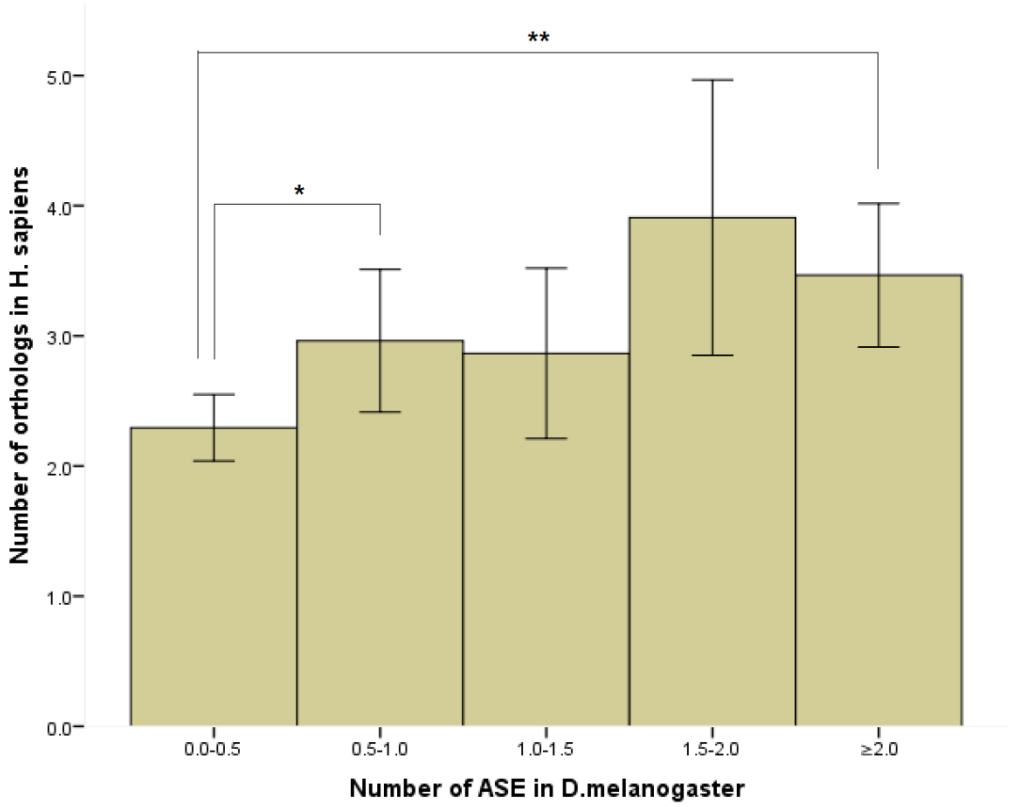
Figure S2 Comparison of number of orthologous between two gene categories (with and without ASE in *C.elegans*) in all vertebrates species (single copy (A) all genes (B) and multiple copies (C). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.


**
*
4.0
3.0
2.0
1.0
0.0
Number of orthologs in Vertebrates
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in D.melanogaster

A

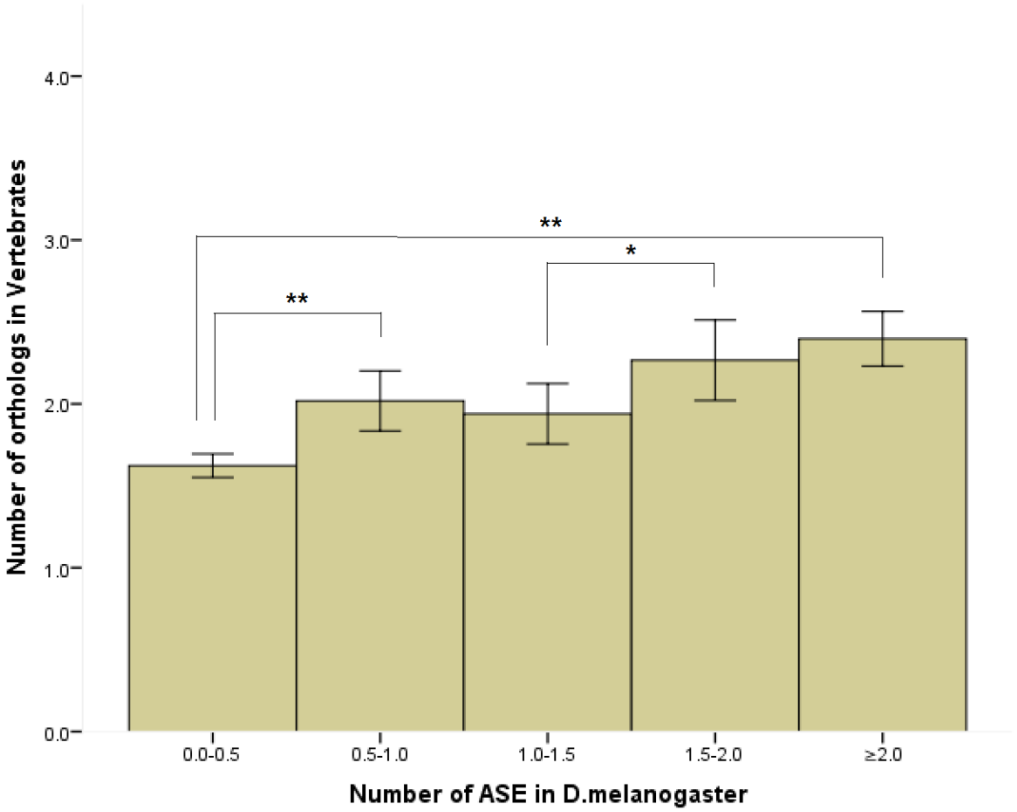

**
*
4.0
3.0
2.0
1.0
0.0
Number of orthologs in G. gallus
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in D.melanogaster

B

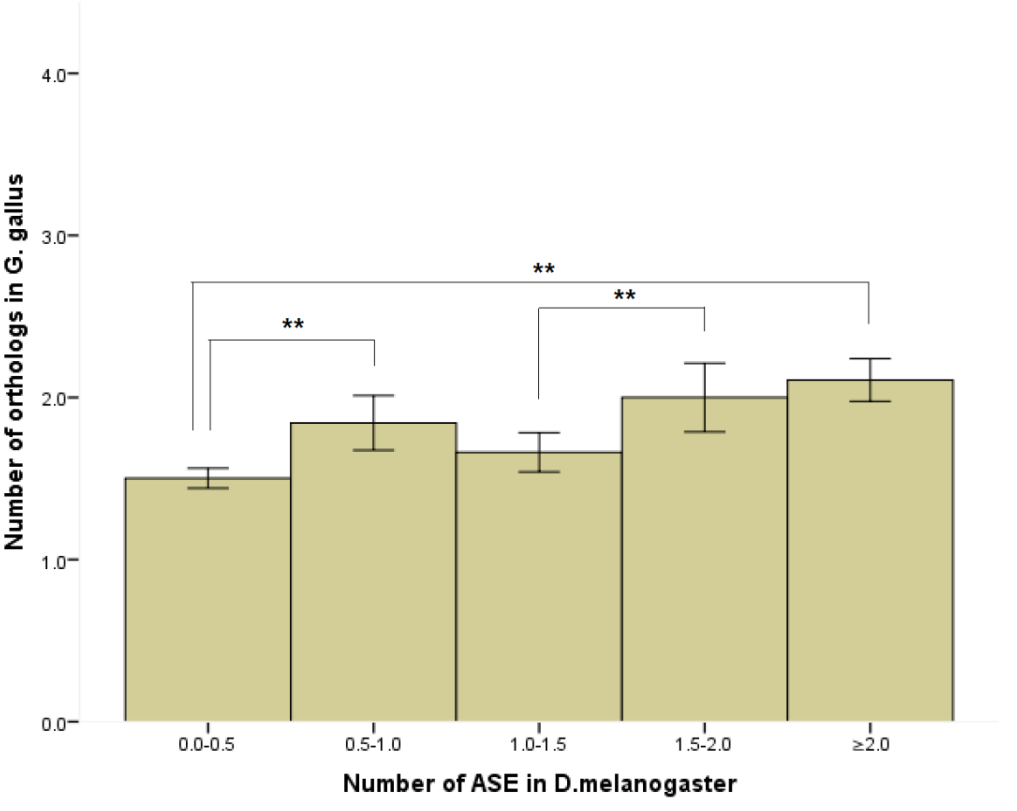


C

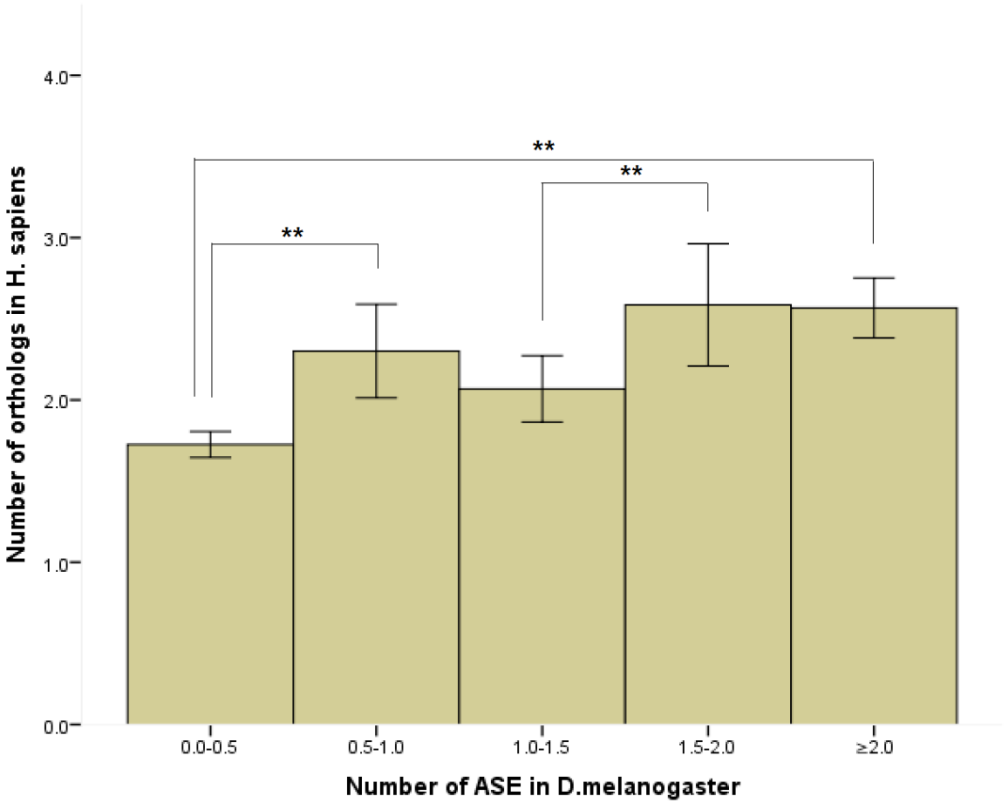
Figure S3 Comparison between different ASEs level of multi-copy genes in *D. melanogaster* and number of orthologous in vertebrates. (A) Average of vertebrates without *D. rerio* (Fish), (B) *G. gallus* (Chicken), (C) *H. sapien* (Human). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.


Number of orthologs in Vertebrates
4.0
3.0
2.0
1.0
0.0
**
*
**
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in D.melanogaster

A

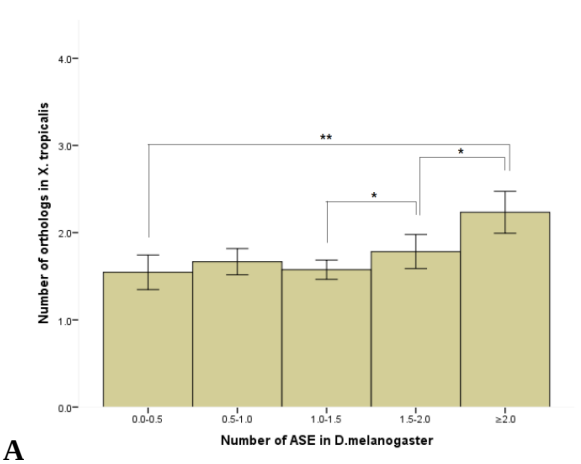

Number of orthologs in G. gallus
4.0
3.0
2.0
1.0
0.0
**
**
**
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in D.melanogaster

B

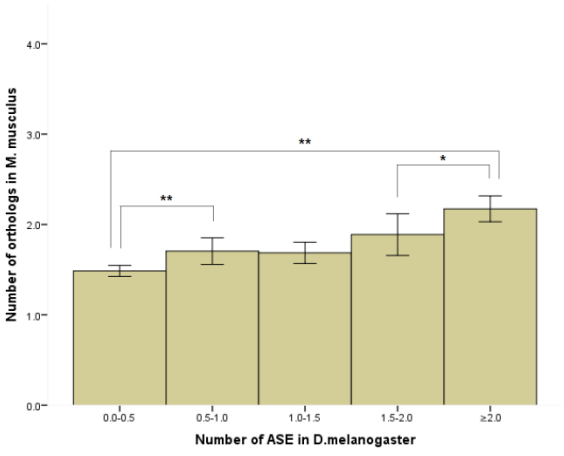


C

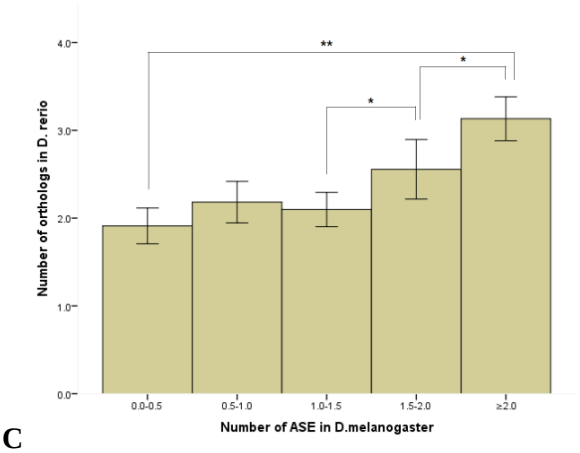
Figure S4 Comparison between different ASEs level of all genes in *D. melanogaster* and number of orthologous in vertebrates. (A) Average of vertebrates without *D. rerio* (Fish), (B) *G. gallus* (Chicken), (C) *H. sapien* (Human). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.



A

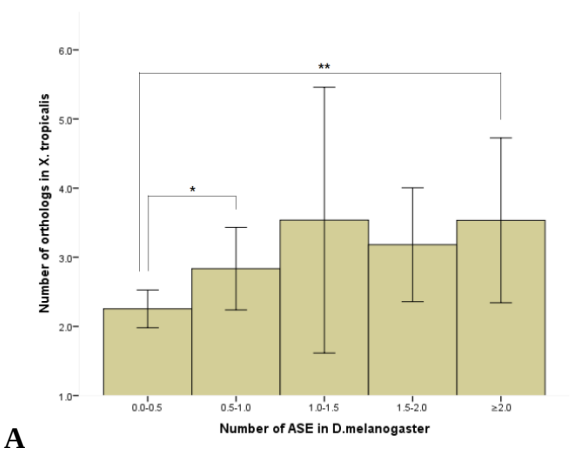


B

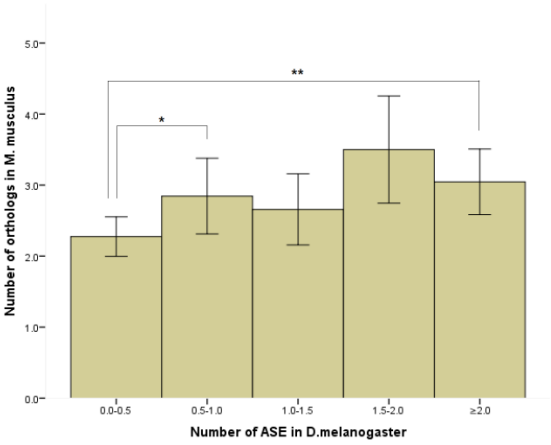


C

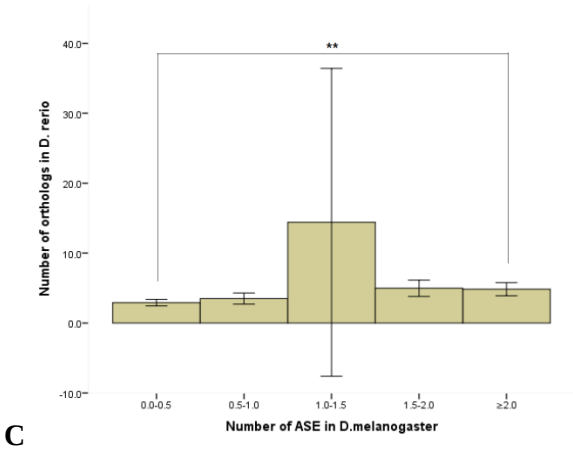
Figure S5 Comparison between different ASEs level of single copy genes in *D. melanogaster* and number of orthologous in vertebrates. (A) *X. tropicalis* (Frog), (B) *M. musculus* (Mice), (C) *D. rerio* (Fish). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.



A



B



C

Figure S6 Comparison between different ASEs level of multi copy genes in *D. melanogaster* and number of orthologous in vertebrates. (A) *X. tropicalis* (Frog), (B) *M. musculus* (Mice), (C) *D. rerio* (Fish). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.

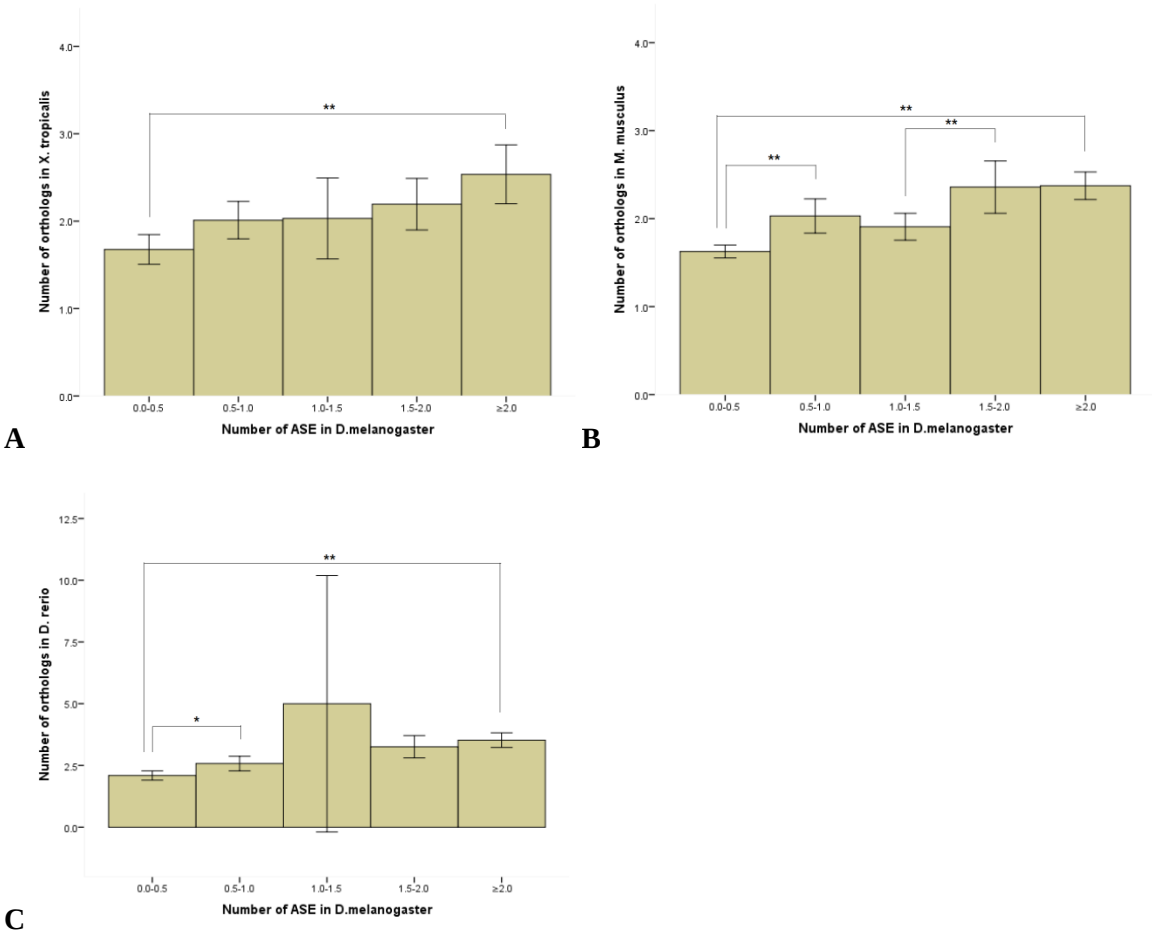
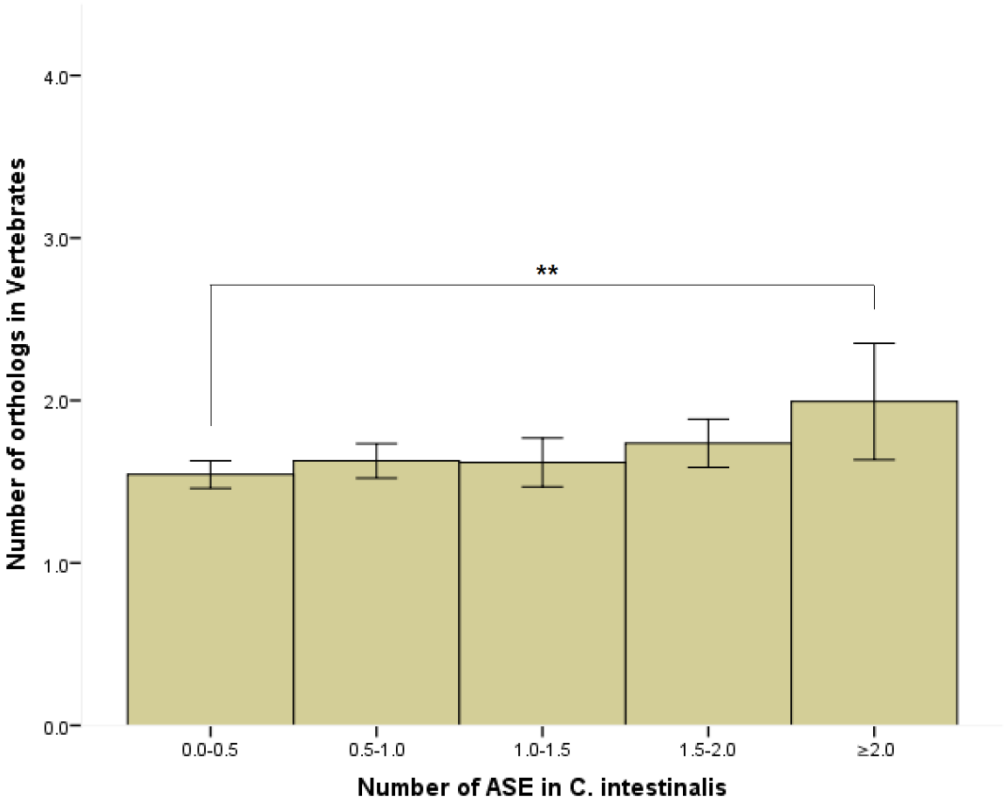
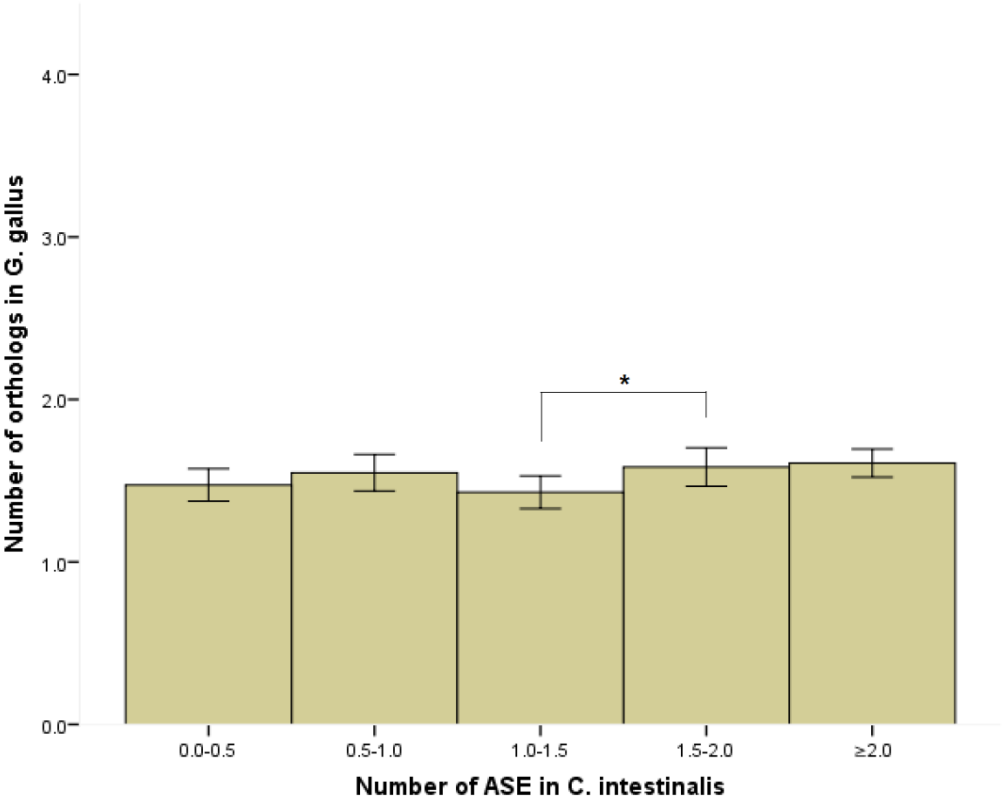


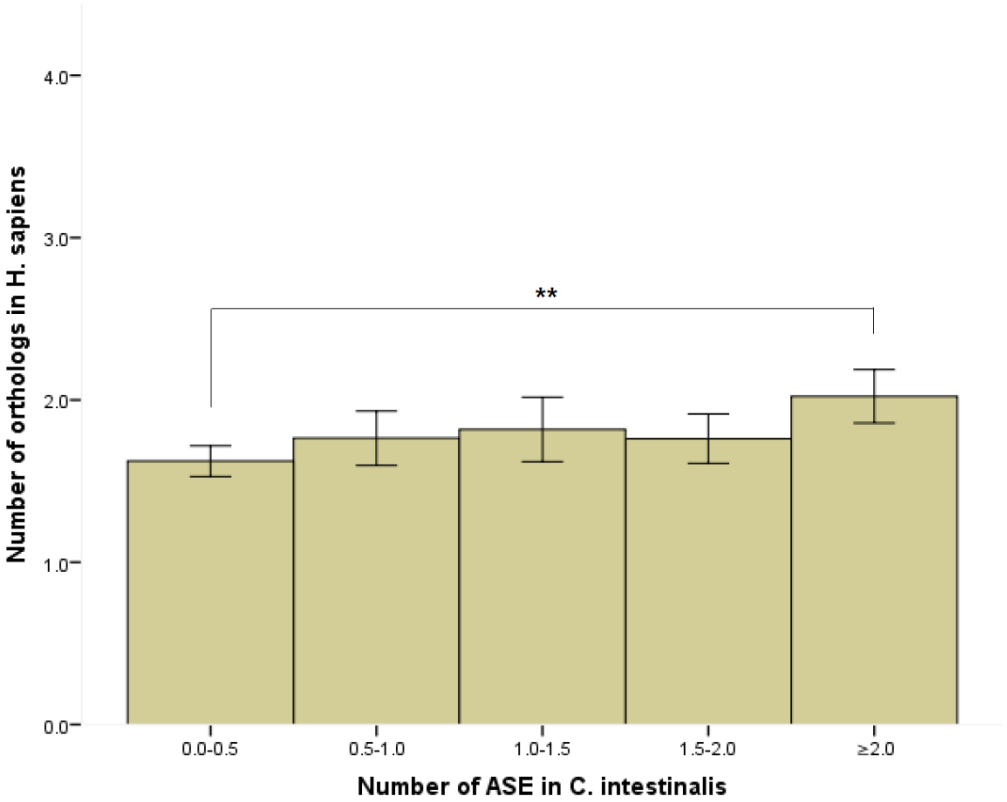
Figure S7 Comparison between different ASEs level of all genes in *D. melanogaster* and number of orthologous in vertebrates. (A) *X. tropicalis* (Frog), (B) *M. musculus* (Mice), (C) *D. rerio* (Fish). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.


Number of orthologs in Vertebrates
4.0
3.0
2.0
1.0
0.0
**
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

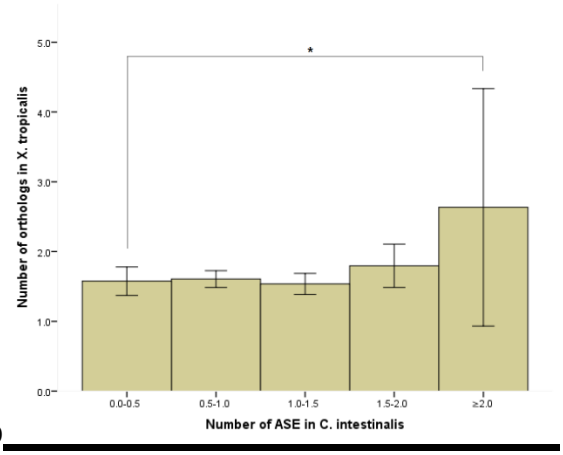
A


Number of orthologs in G. gallus
4.0
3.0
2.0
1.0
0.0
*
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

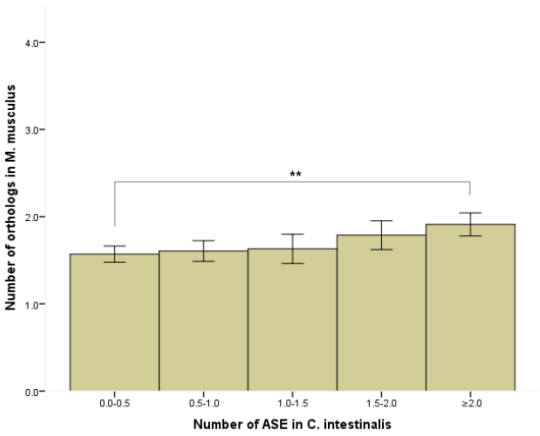
B


Number of orthologs in H. sapiens
4.0
3.0
2.0
1.0
0.0
**
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

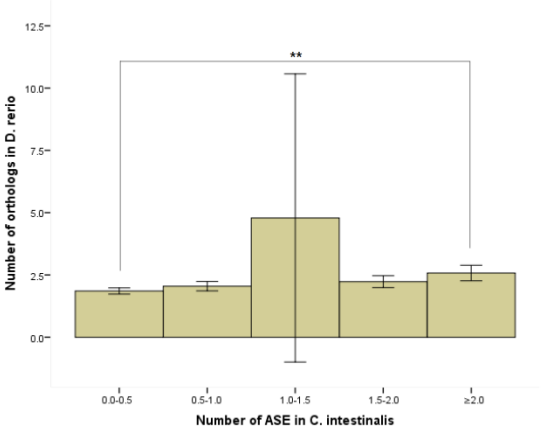
C


Number of orthologs in X. tropicalis
5.0
4.0
3.0
2.0
1.0
0.0
*
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

D

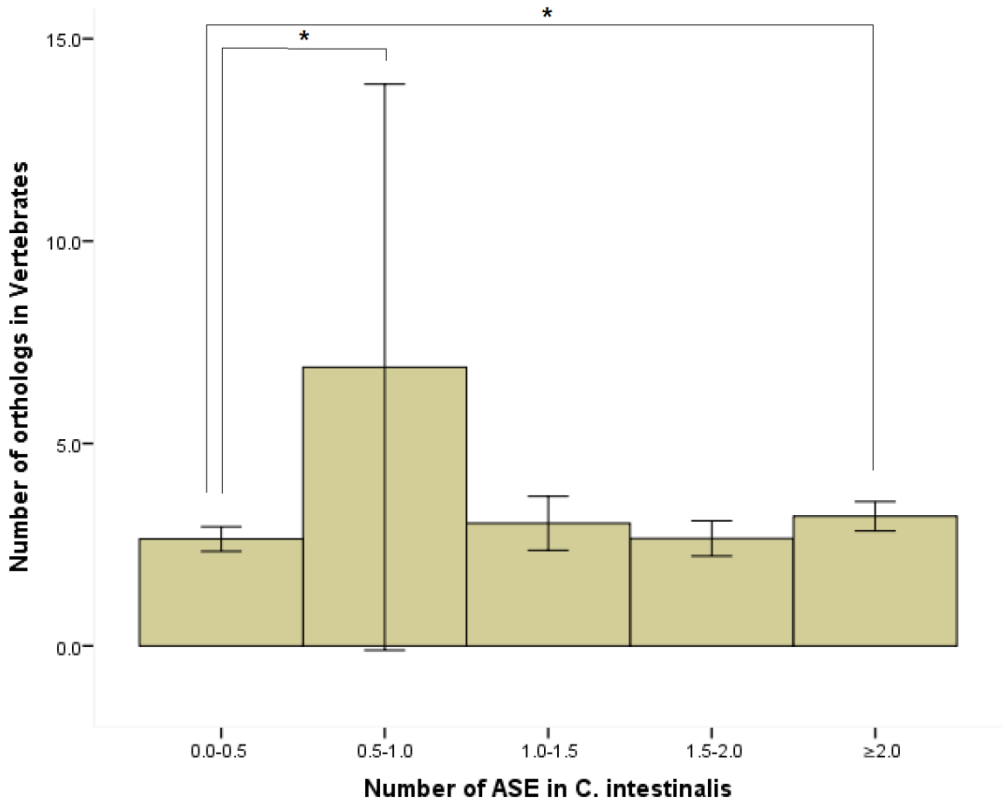

Number of orthologs in M. musculus
4.0
3.0
2.0
1.0
0.0
**
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

E

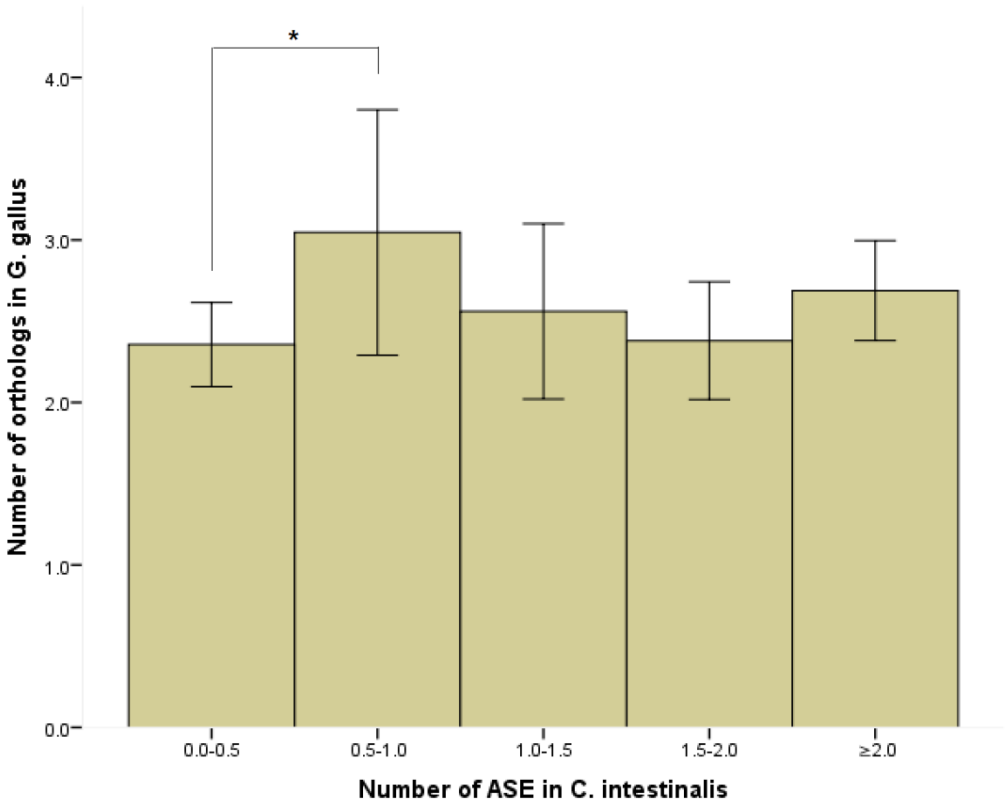

Number of orthologs in D. rerio
12.5
10.0
7.5
5.0
2.5
0.0
**
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

F

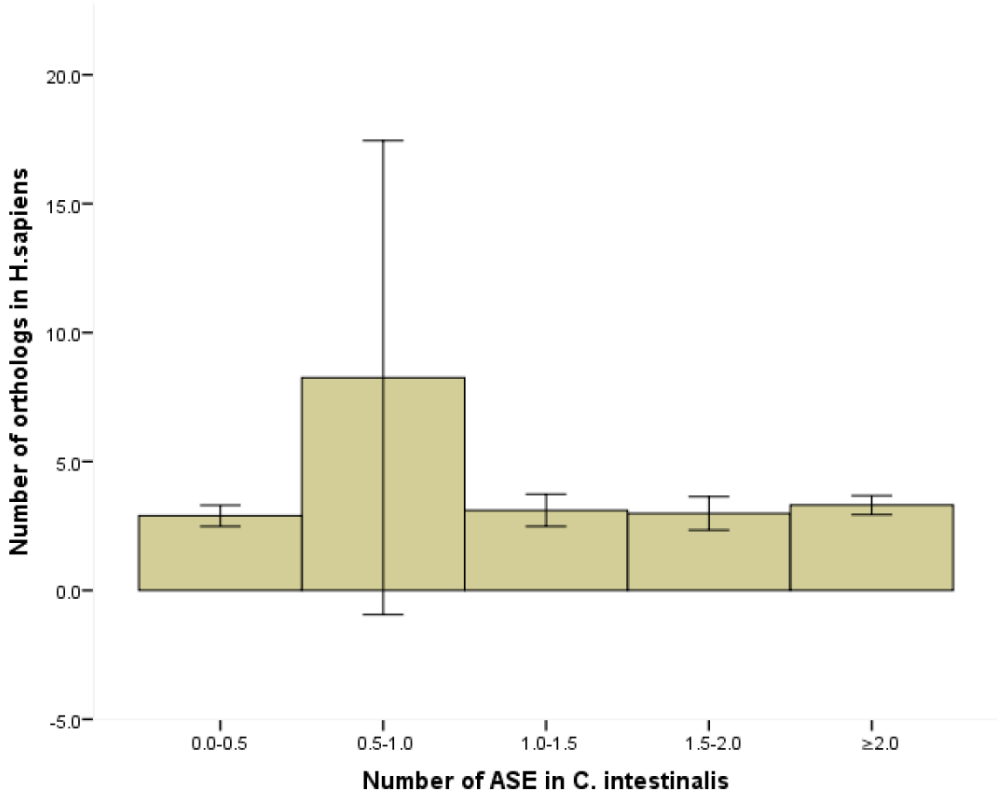
Figure S8 Comparison between different ASEs level of single copy genes in *C.intestinalls* and number of orthologous in vertebrates. (A) Average of vertebrates without *D. rerio* (Fish), (B) *G. gallus* (Chicken), (C) *H. sapien* (Human). (D) *X. tropicalis* (Frog), (E) *M. musculus* (Mice), (F) *D. rerio* (Fish). *($P<0.05$) means significantly different *($P<0.05$) means significantly different, **($P<0.01$) shows very significant difference.


*
*
15.0
10.0
5.0
0.0
Number of orthologs in Vertebrates
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

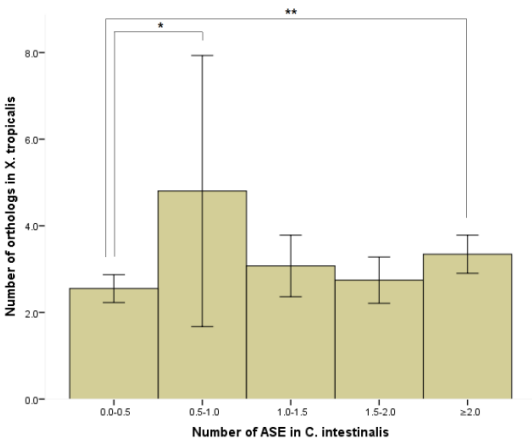
A


*
4.0
3.0
2.0
1.0
0.0
Number of orthologs in G. gallus
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

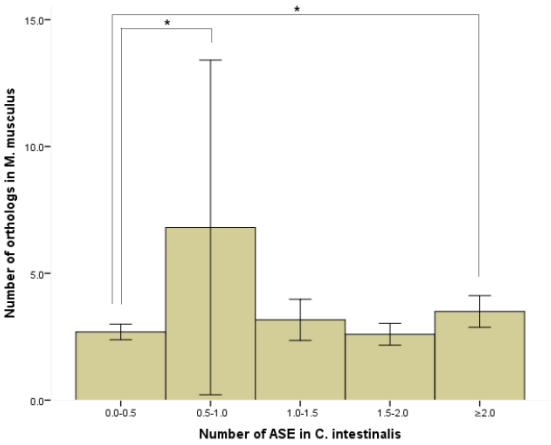
B


Number of orthologs in H.sapiens
20.0
15.0
10.0
5.0
0.0
-5.0
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

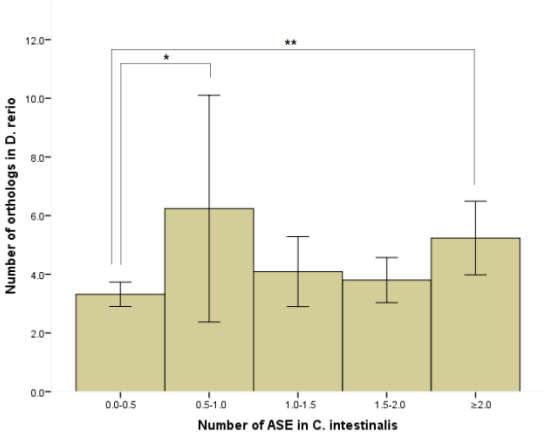
C


Number of orthologs in X. tropicalis
*
**
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

D

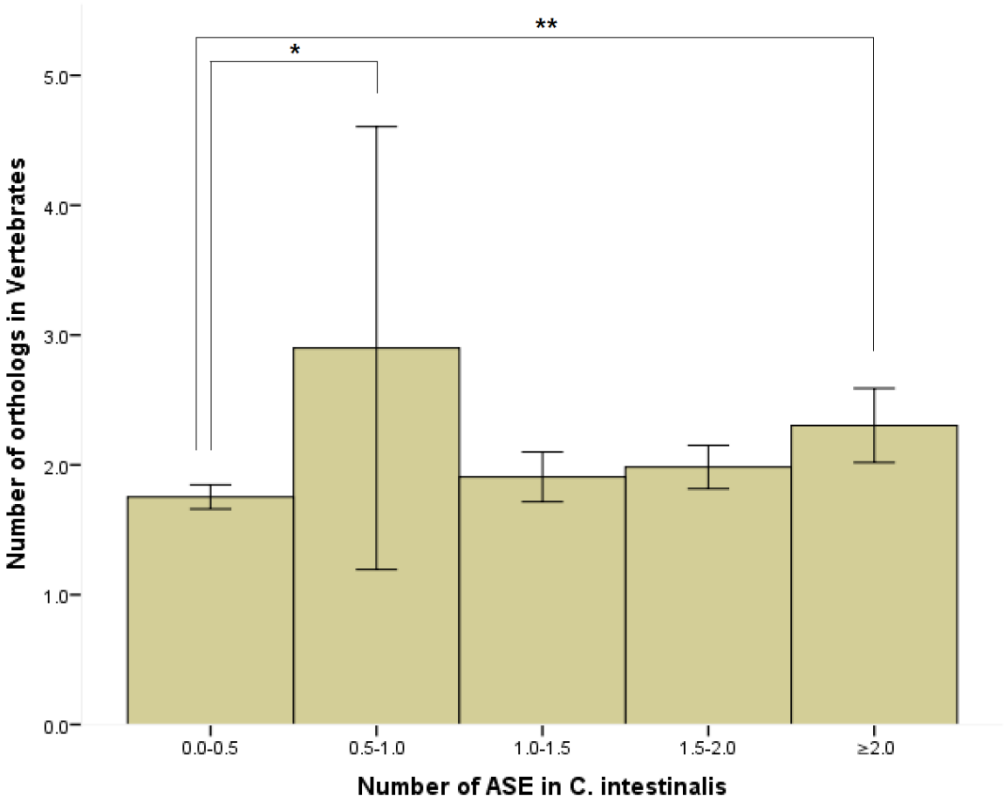

Number of orthologs in M. musculus
*
*
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

E

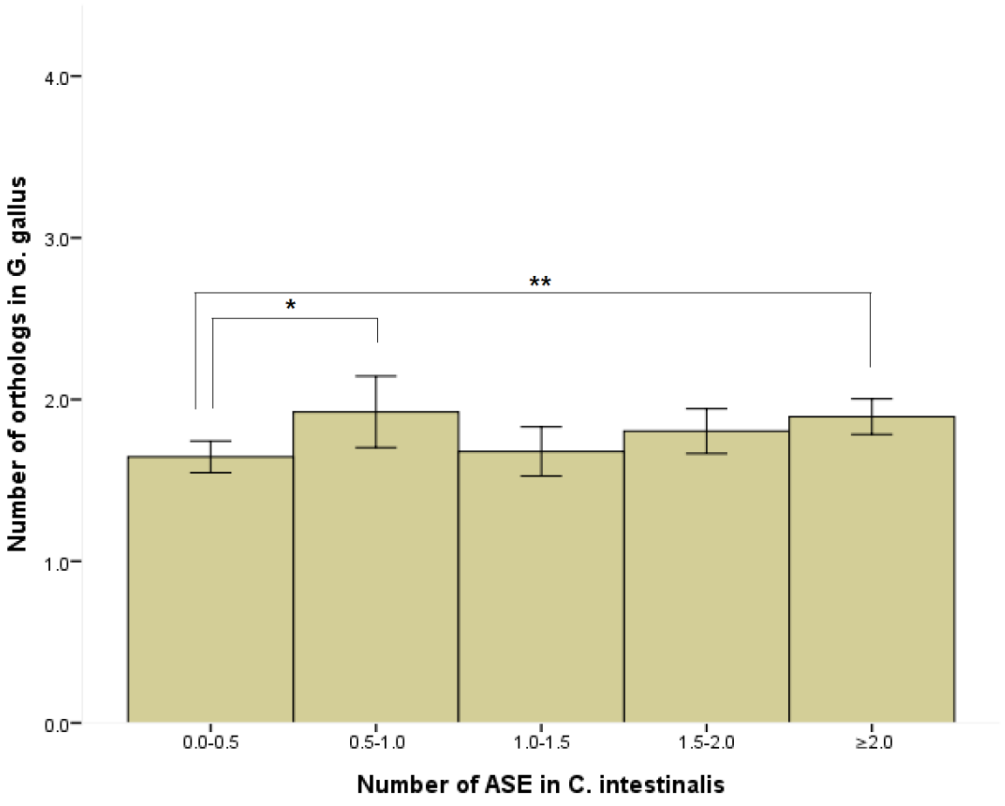

Number of orthologs in D. rerio
*
**
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

F

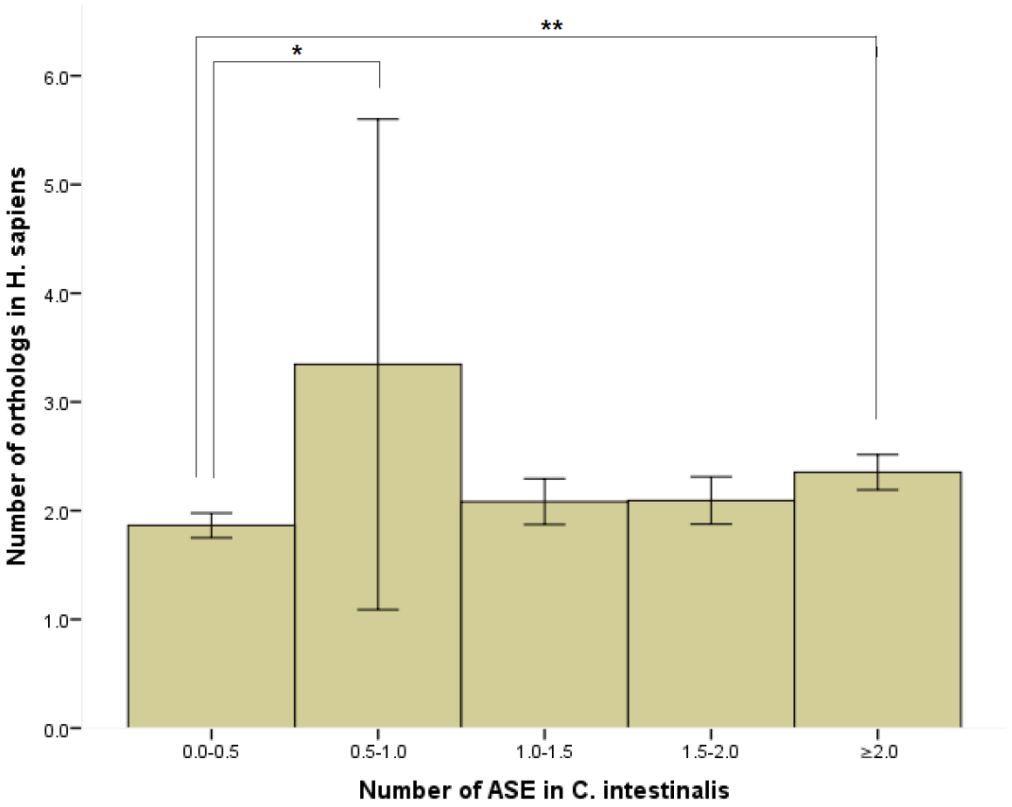
Figure S9 Comparison between different ASEs level of multi copy genes in *C.intestinalls* and number of orthologous in vertebrates. (A) average of vertebrates without *D. rerio* (Fish), (B) *G. gallus* (Chicken), (C) *H. sapien* (Human). (D) *X. tropicalis* (Frog), (E) *M. musculus* (Mice), (F) *D. rerio* (Fish). *($P<0.05$) means significantly different *($P<0.05$) means significantly different, **($P<0.01$) shows very significant difference.


**
*
5.0
4.0
3.0
2.0
1.0
0.0
Number of orthologs in Vertebrates
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

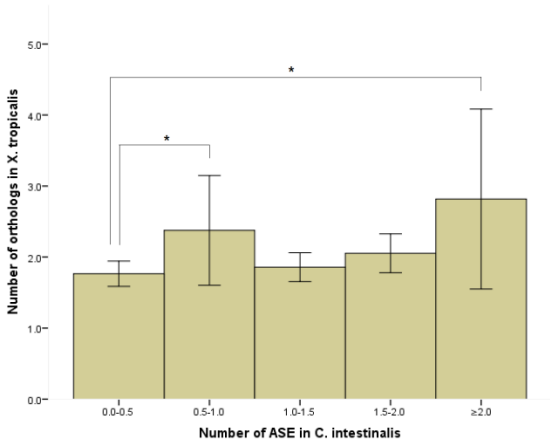
A


**
*
4.0
3.0
2.0
1.0
0.0
Number of orthologs in G. gallus
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

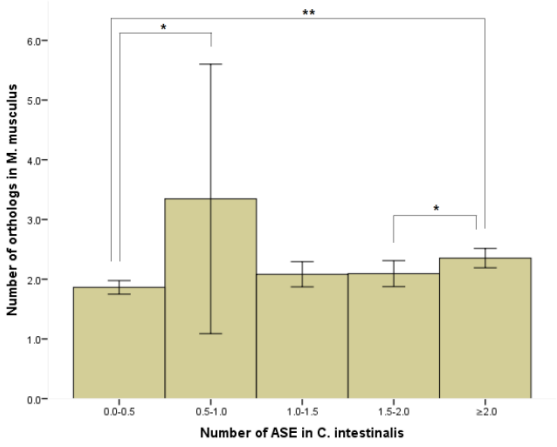
B


**
*
6.0
5.0
4.0
3.0
2.0
1.0
0.0
Number of orthologs in H. sapiens
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

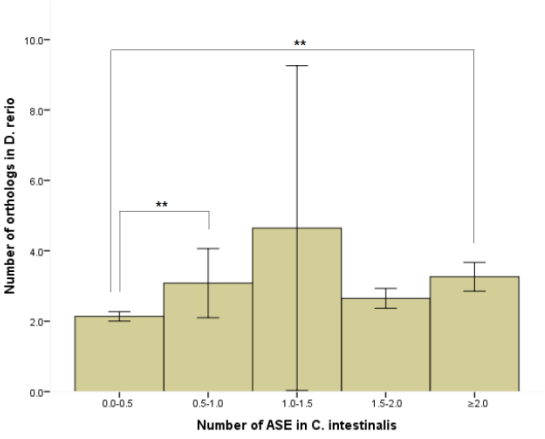
C


*
*
5.0
4.0
3.0
2.0
1.0
0.0
Number of orthologs in X. tropicalis
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

D

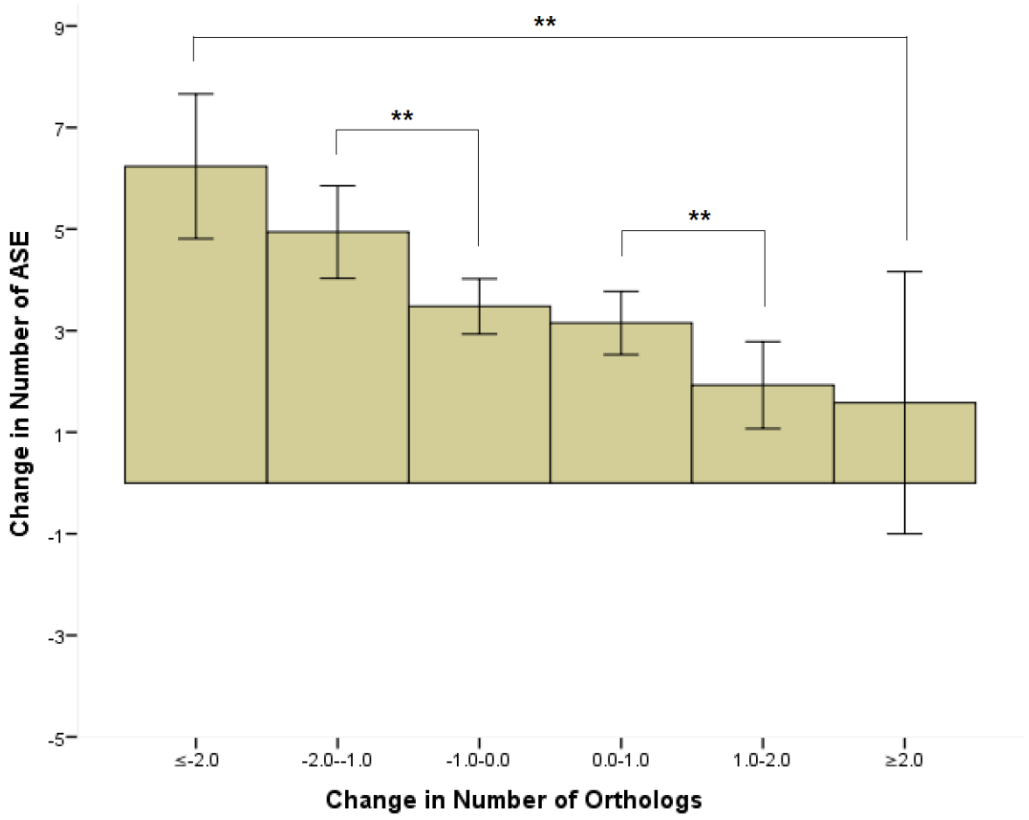

**
*
*
6.0
5.0
4.0
3.0
2.0
1.0
0.0
Number of orthologs in M. musculus
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

E

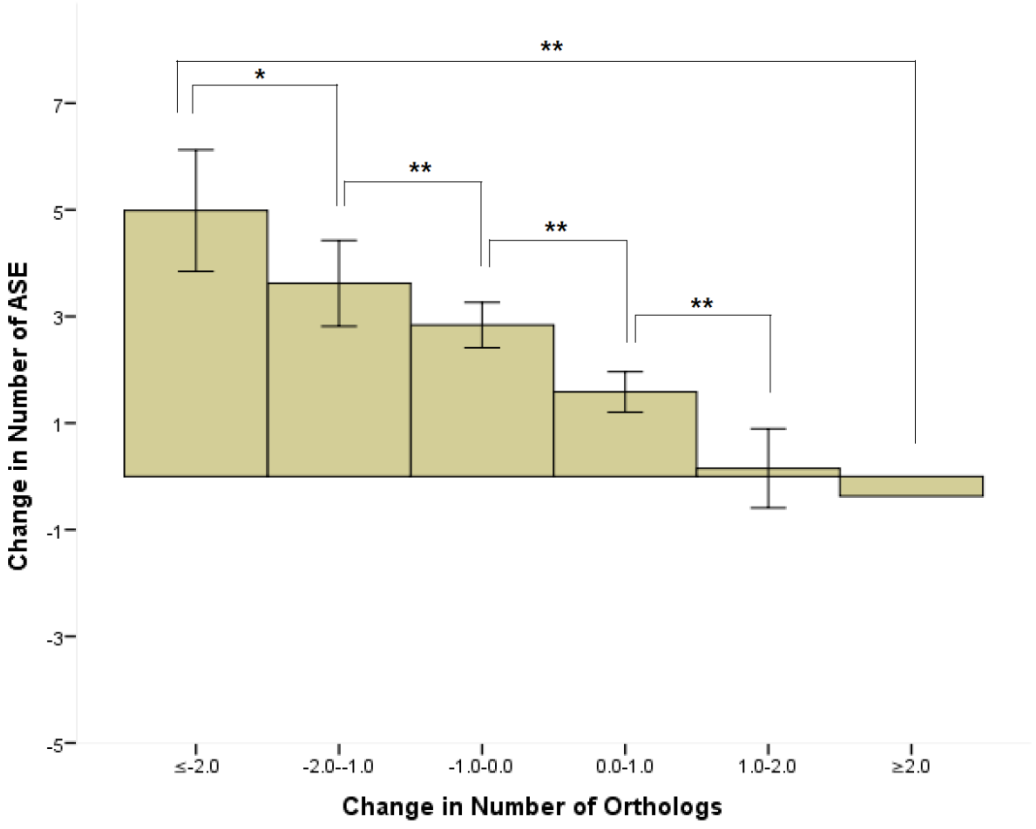

**
**
10.0
8.0
6.0
4.0
2.0
0.0
Number of orthologs in D. rerio
0.0-0.5
0.5-1.0
1.0-1.5
1.5-2.0
≥2.0
Number of ASE in C. intestinalis

F

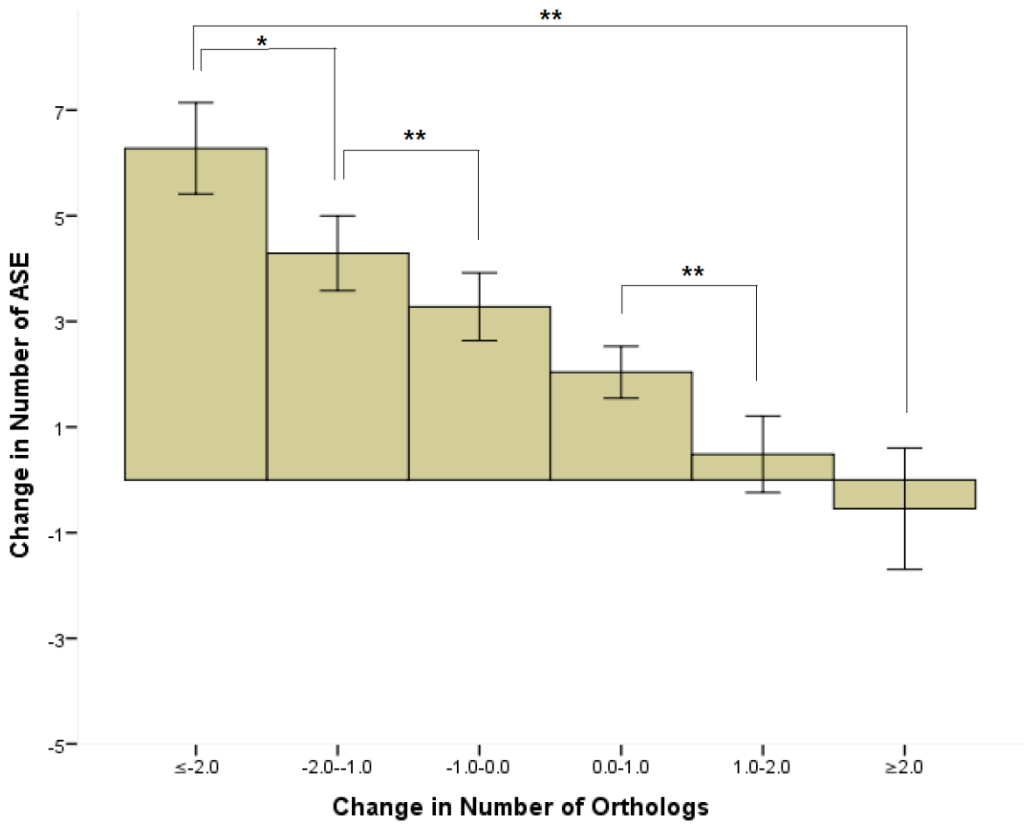
Figure S10 Comparison between different ASEs level of all genes in *C.intestinalls* and number of orthologous in vertebrates. (A) average of vertebrates without *D. rerio* (Fish), (B) *G. gallus* (Chicken), (C) *H. sapien* (Human). (D) *X. tropicalis* (Frog), (E) *M. musculus* (Mice), (F) *D. rerio* (Fish). *($P<0.05$) means significantly different *($P<0.05$) means significantly different, **($P<0.01$) shows very significant difference.


**
**
**
9
7
5
3
1
-1
-3
-5
Change in Number of ASE
≤-2.0
-2.0--1.0
-1.0-0.0
0.0-1.0
1.0-2.0
≥2.0
Change in Number of Orthologs

A

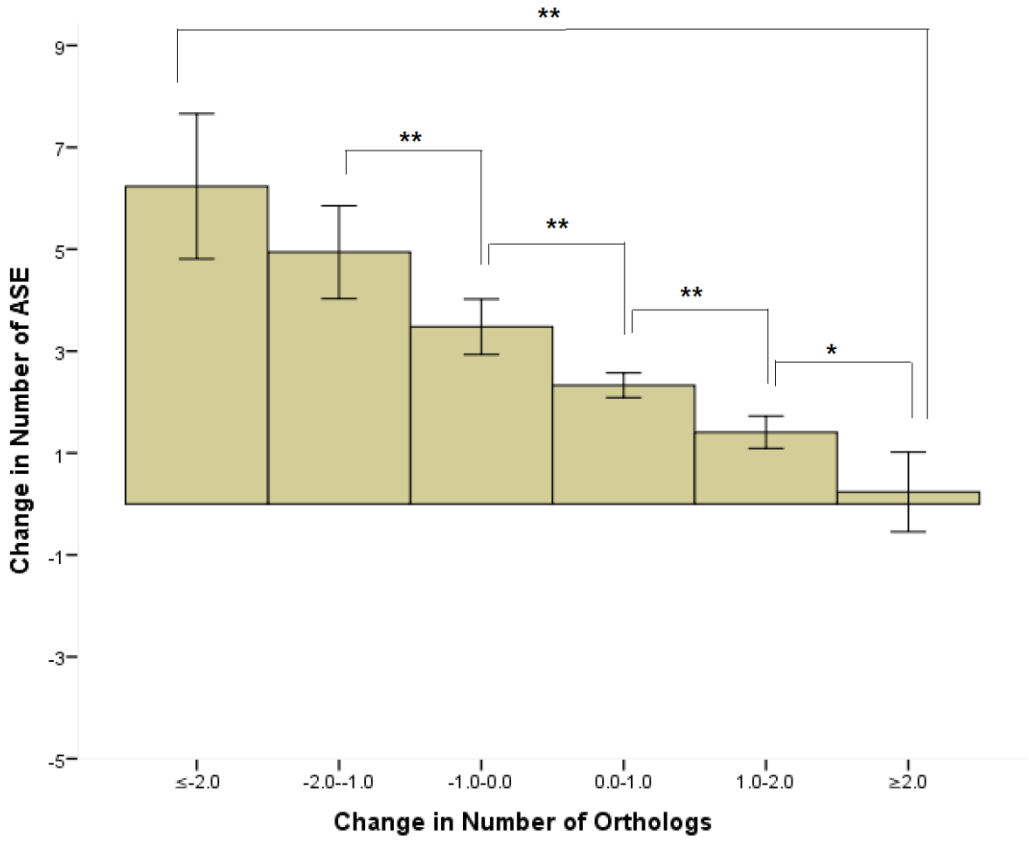

**
*
**
**
**
7
5
3
1
-1
-3
-5
Change in Number of ASE
≤-2.0
-2.0--1.0
-1.0-0.0
0.0-1.0
1.0-2.0
≥2.0
Change in Number of Orthologs

B

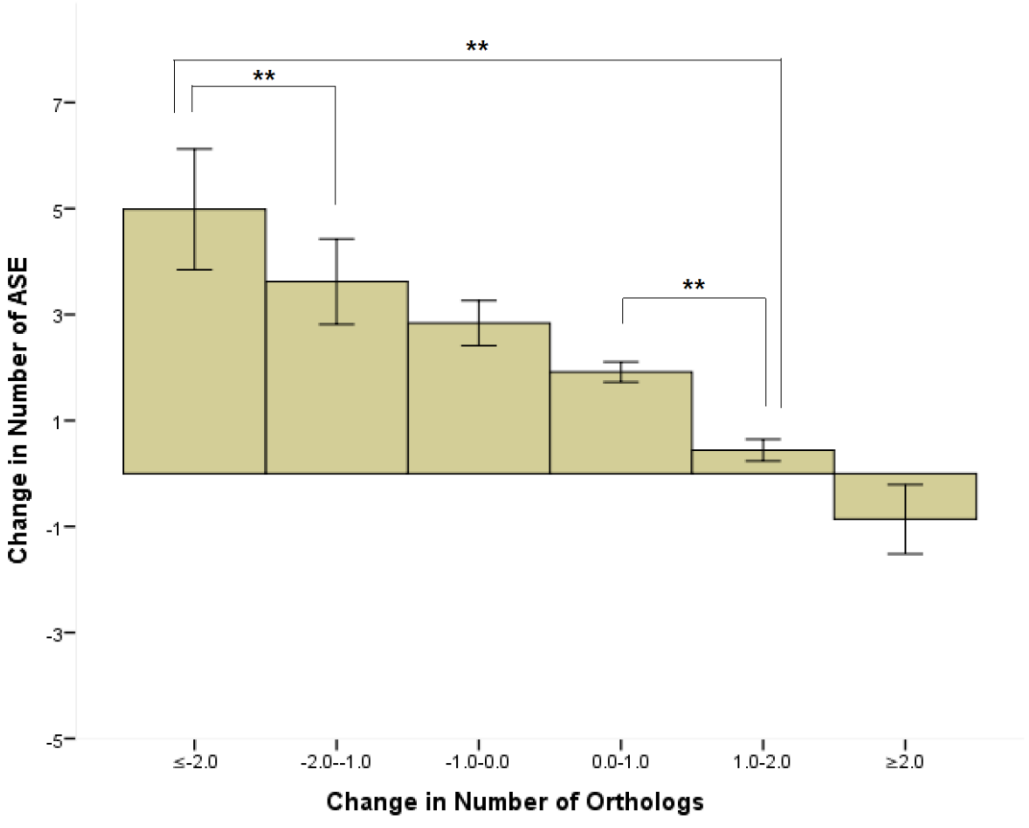


C

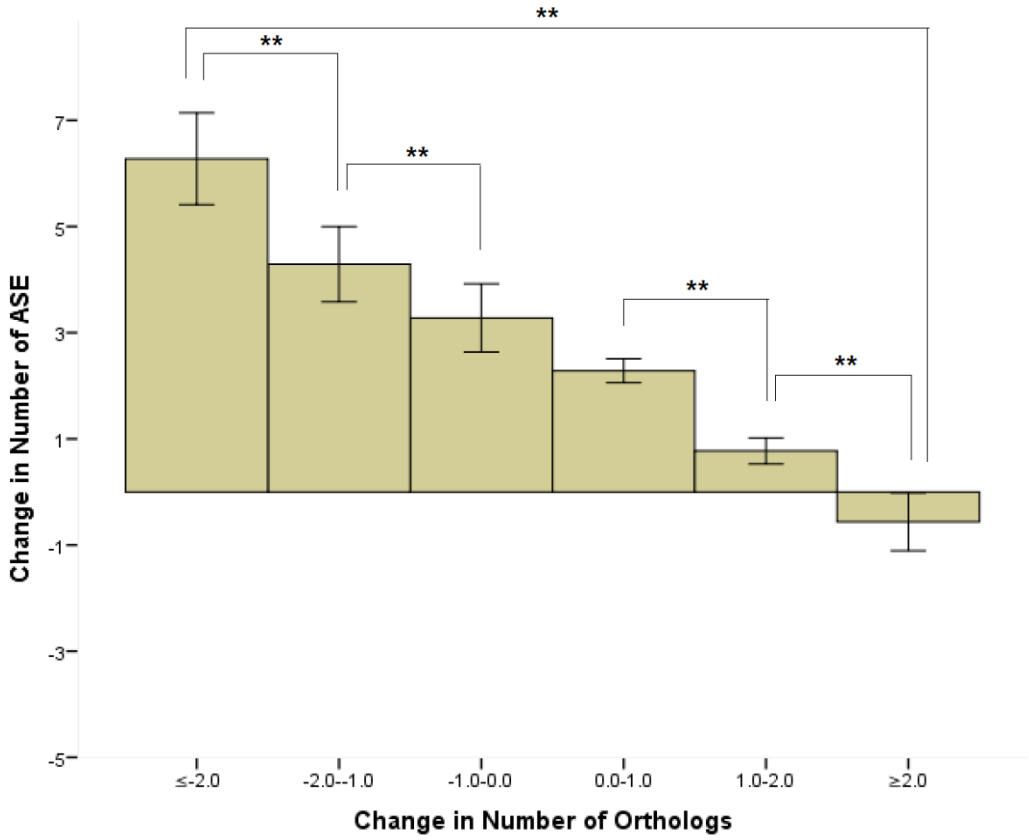
Figure S11 Change in Number of orthologous and Number of ASE between *D. melanogaster* and vertebrates. Multi-copy genes in fly (A) average of vertebrates without *D. rerio* (Fish), (B) *G. gallus* (Chicken), (C) *H. sapien* (Human). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.


**
**
**
**
*
9
7
5
3
1
-1
-3
-5
Change in Number of ASE
≤-2.0
-2.0--1.0
-1.0-0.0
0.0-1.0
1.0-2.0
≥2.0
Change in Number of Orthologs

A


**
**
**
7
5
3
1
-1
-3
-5
Change in Number of ASE
≤-2.0
-2.0--1.0
-1.0-0.0
0.0-1.0
1.0-2.0
≥2.0
Change in Number of Orthologs

B



C

Figure S12 Change in Number of orthologous and Number of ASE between *D. melanogaster* and vertebrates. all genes in fly (A) average of vertebrates without *D. rerio* (Fish), (B) *G. gallus* (Chicken), (C) *H. sapien* (Human). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.

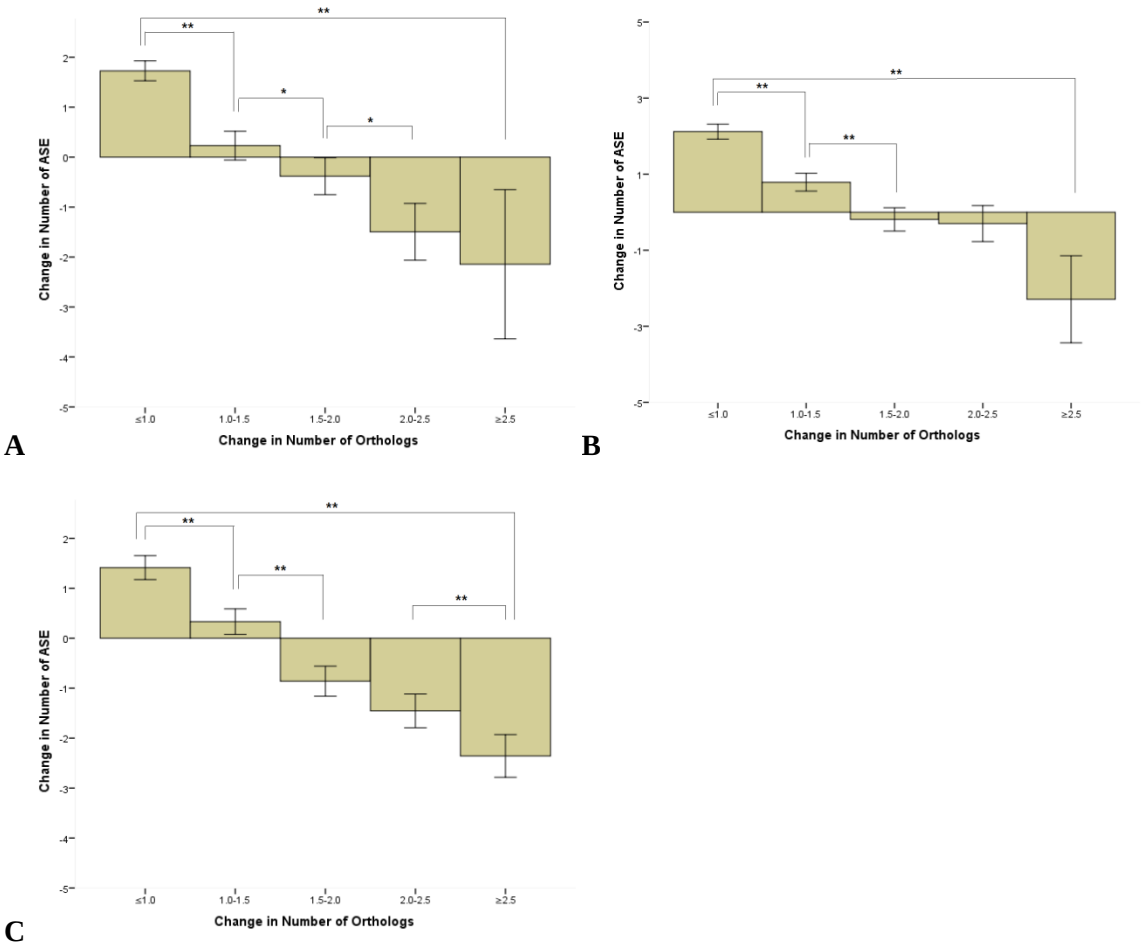


Figure S13 Change in Number of orthologous and Number of ASE between *D. melanogaster* and vertebrates. single copy genes in fly, (A) *X. tropicalis* (Frog), (B) *M. musculus* (Mice), (C) *D. rerio* (Fish). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.

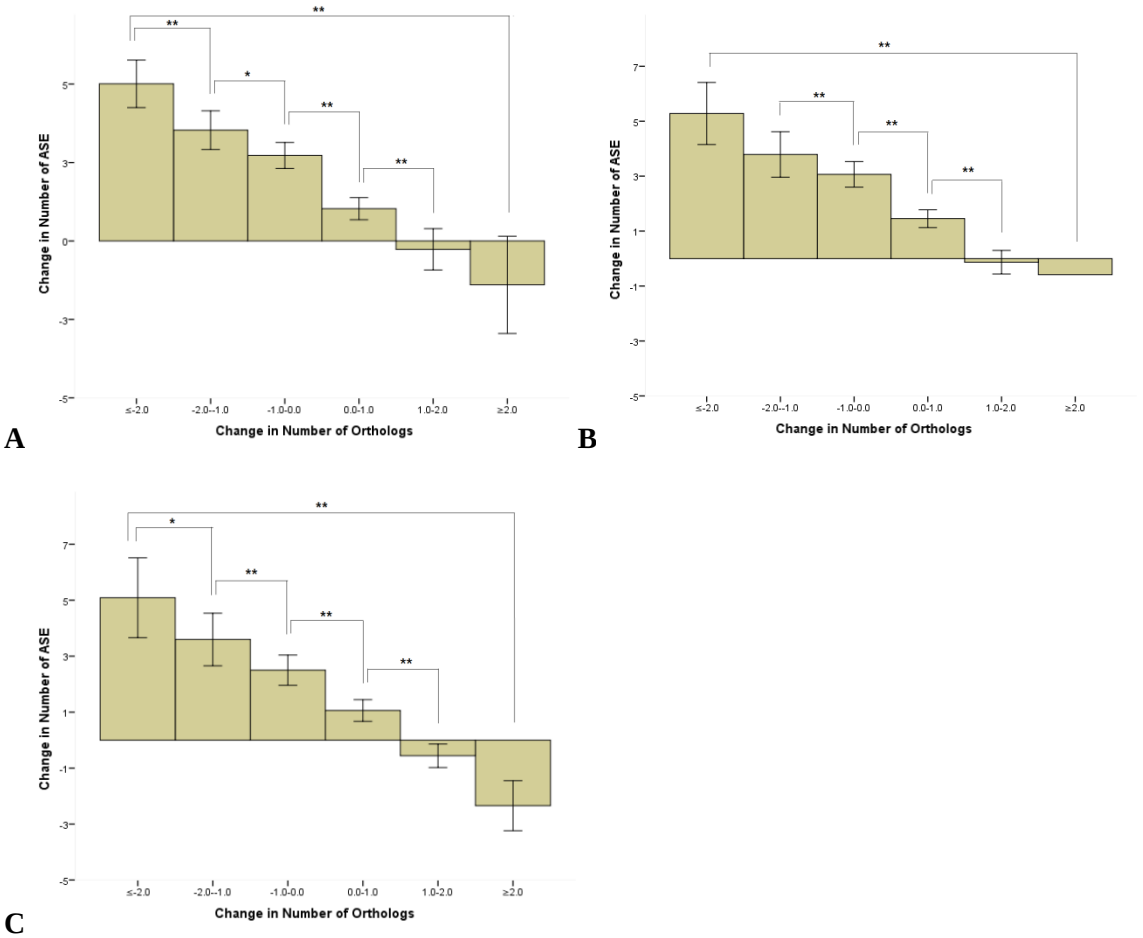


Figure S14 Change in Number of orthologous and Number of ASE between *D. melanogaster* and vertebrates. multi copy genes in fly (A) *X. tropicalis* (Frog), (B) *M. musculus* (Mice), (C) *D. rerio* (Fish). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.

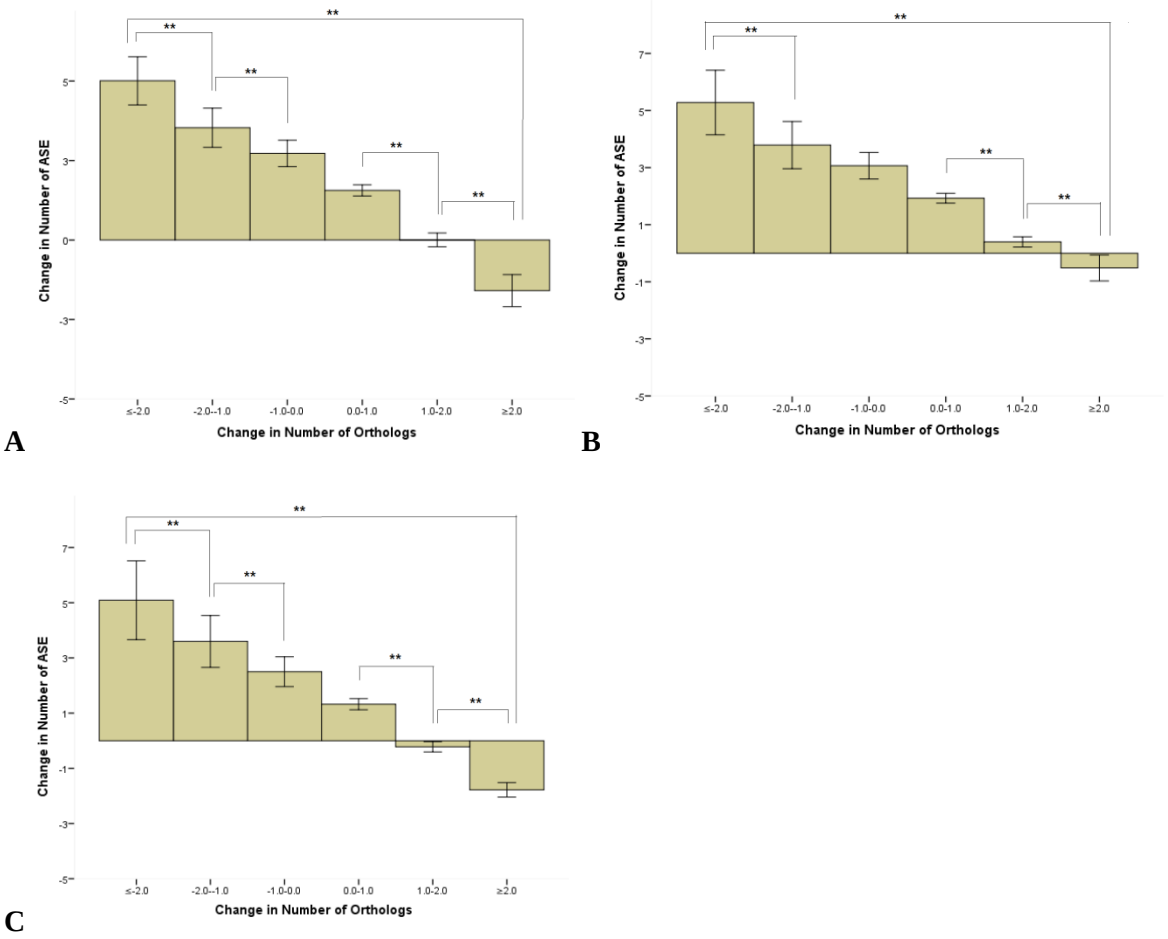


Figure S15 Change in Number of orthologous and Number of ASE between *D. melanogaster* and vertebrates. all genes in fly (A) *X. tropicalis* (Frog), (B) *M. musculus* (Mice), (C) *D. rerio* (Fish). *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.

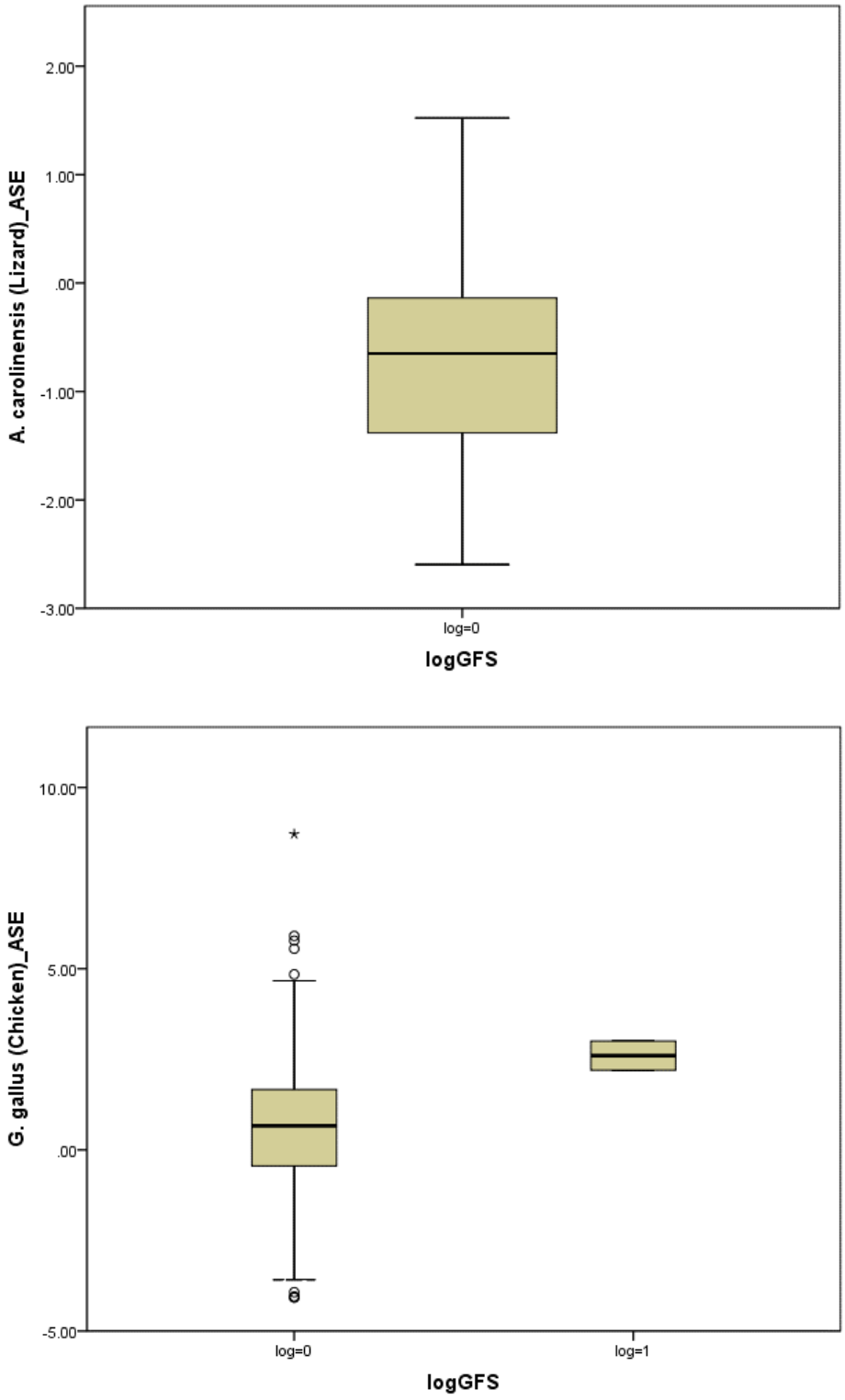


Figure S16 Change in Number of orthologous and Number of ASE between *D. rerio* and other vertebrates. Single copy genes in *D. rerio* compare with (A) *A. carollnensis*, (B) *G. gallus*.

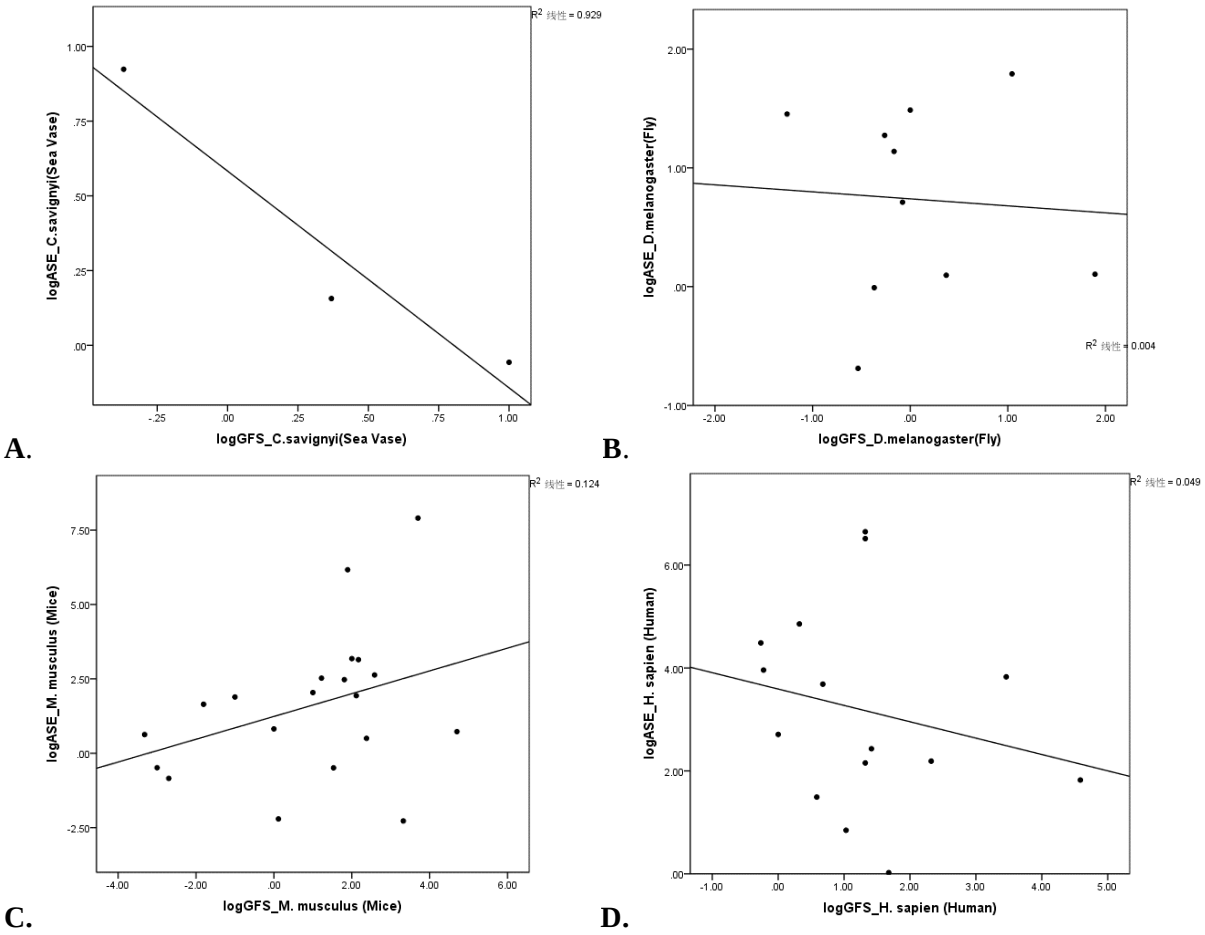


Figure S17 Change in Number of orthologous and Number of ASE between *C.elegan* , invertebrates and vertebrates. Single copy genes in *C.elegan* compare with (A) *C. savignyi*, (B) *D. melanogaster,* (C) *M.musculus*,(D) *H.sapien*.

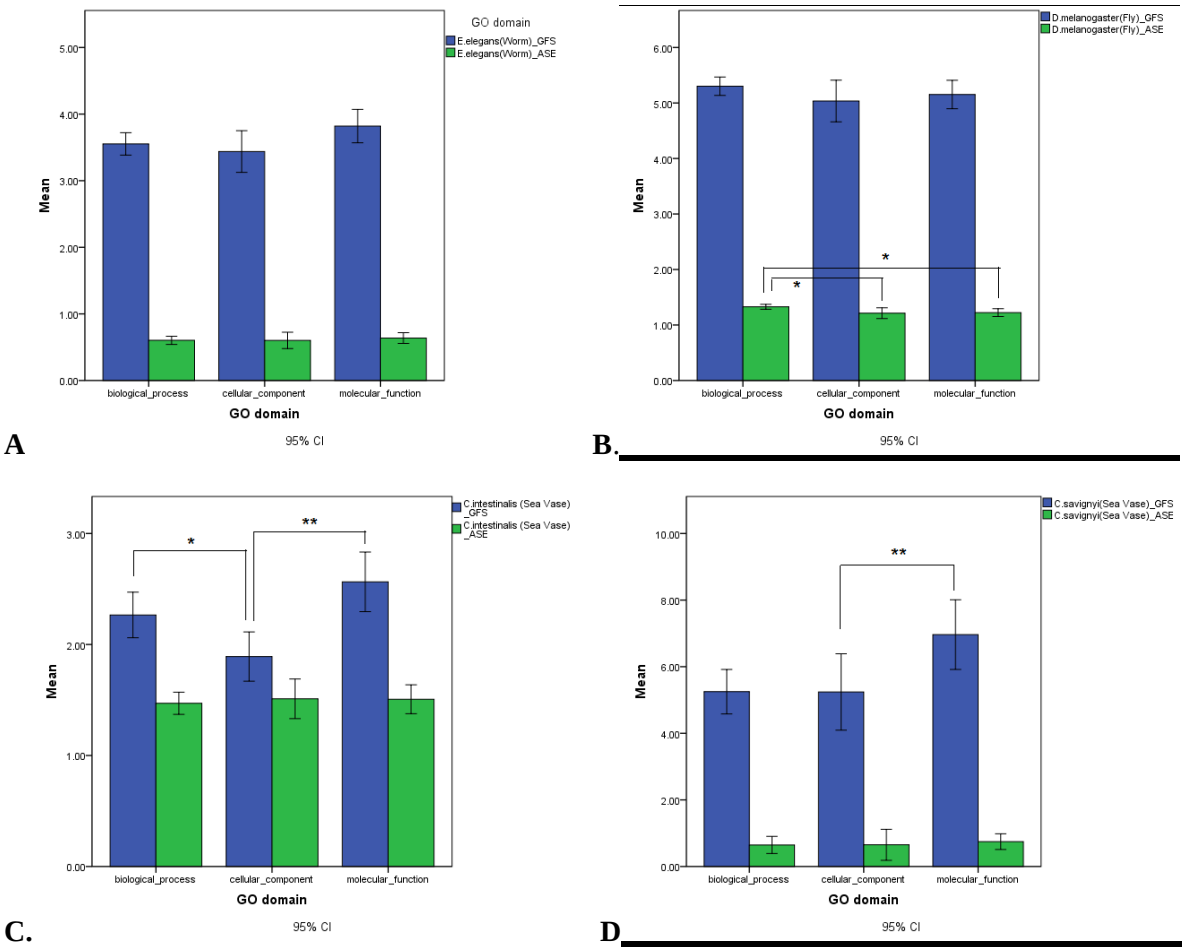


Figure S18 Comparison between the invertebrates GFS and ASE level of three genetic domains (biological process cellular component and molecular function) (A) *E.elegan*, (B) *D.melonogaster*, (C) *C. intestinalis*, (D) *C. savignyi*. *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.

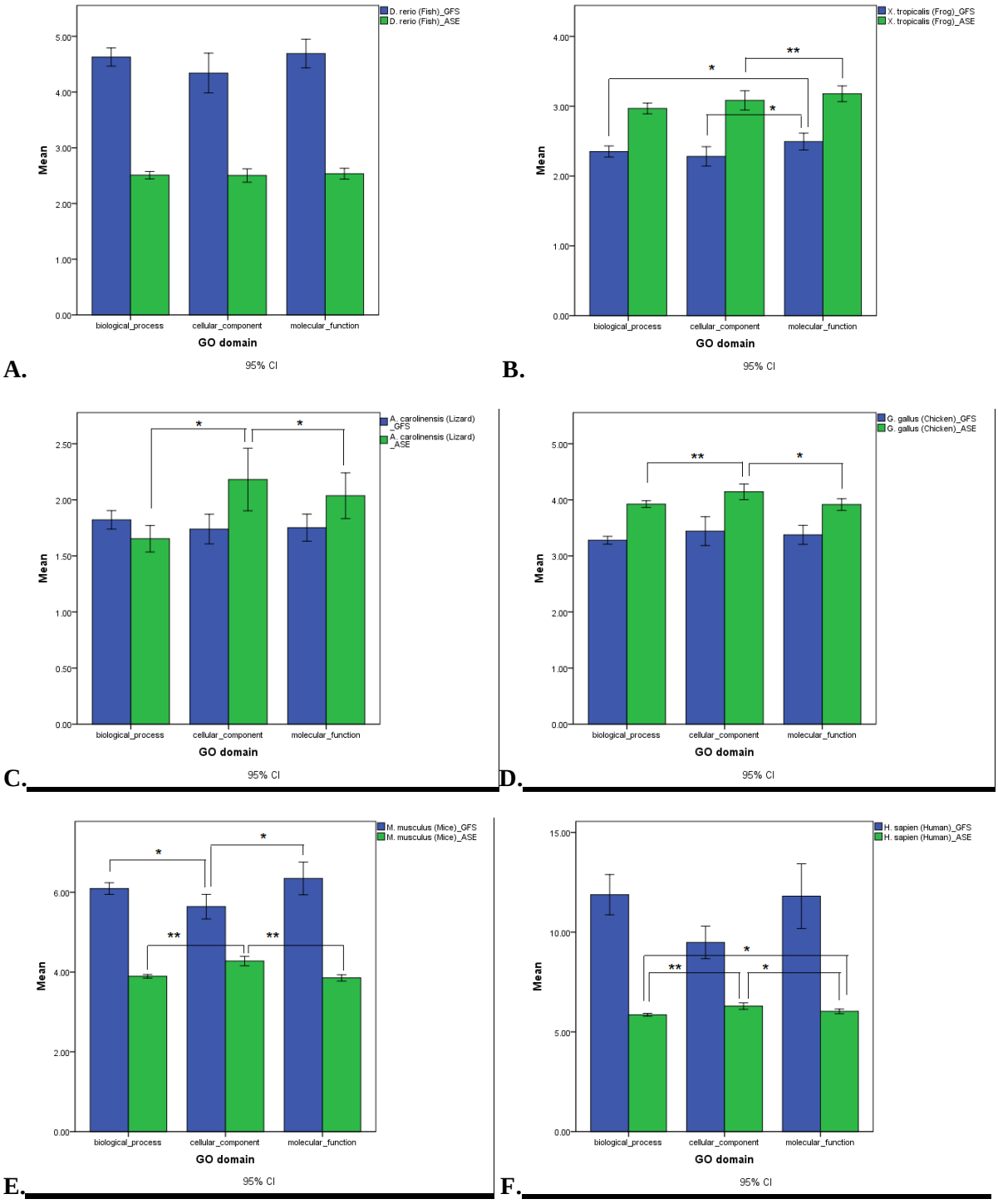


Figure S19 Comparison between the vertebrates GFS and ASE level of three genetic domains (biological process cellular component and molecular function) (A) *D.rerio*, (B) *X. tropicalis* ,(C) *A. carolinensis*, (D) *G.gallus*, (E) *M.musculus*, (F) *H.sapien*, *(P<0.05) means significantly different, **(P<0.01) shows very significant difference.

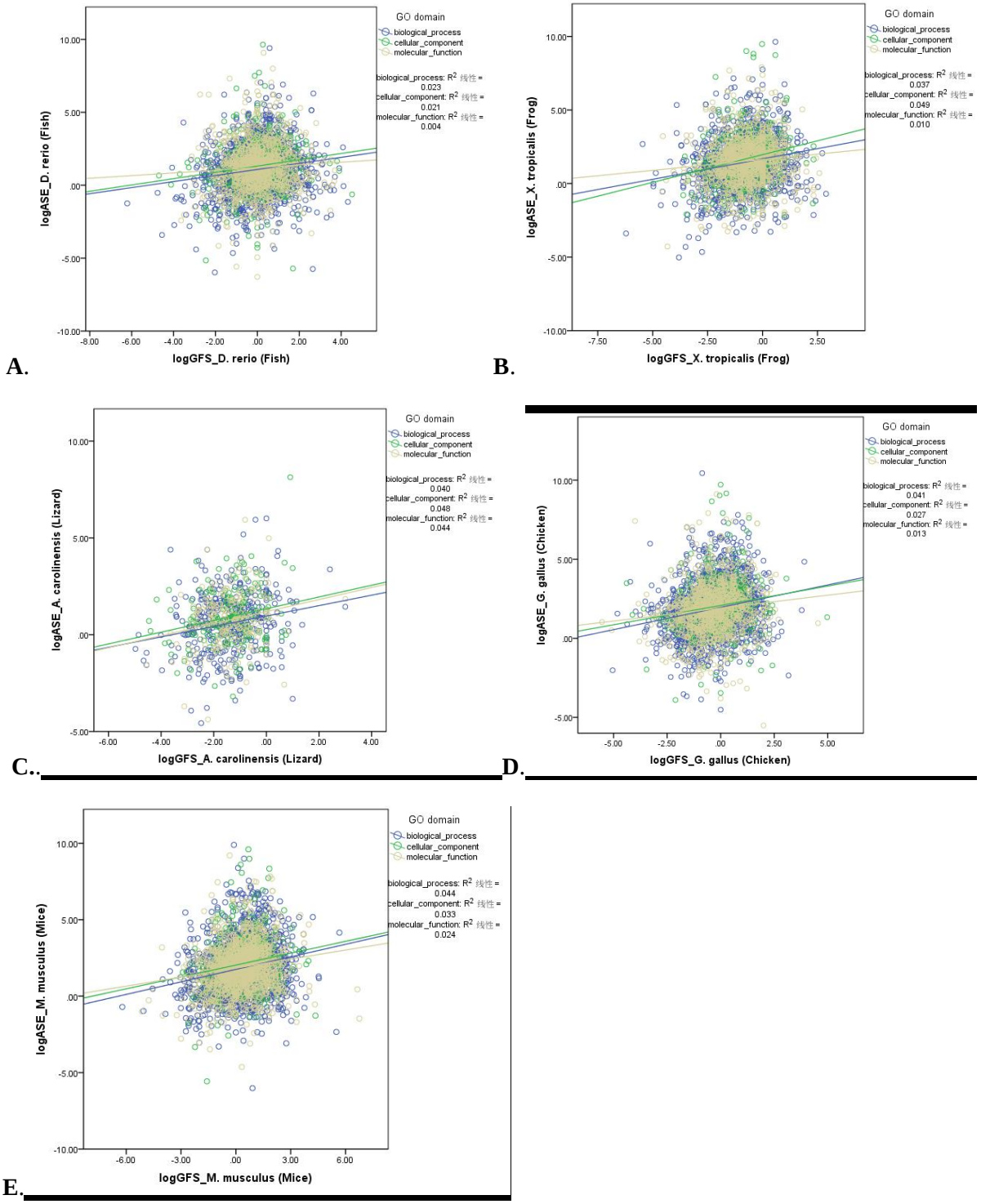
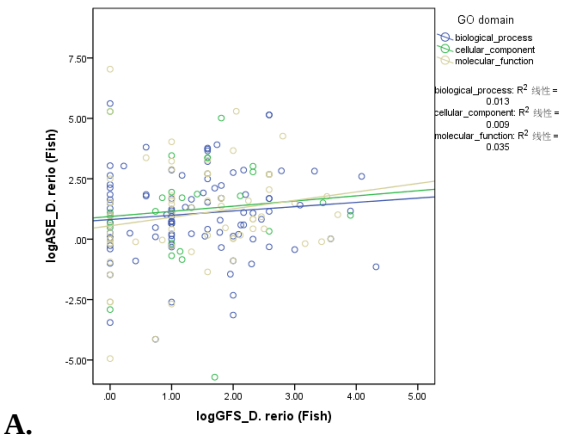
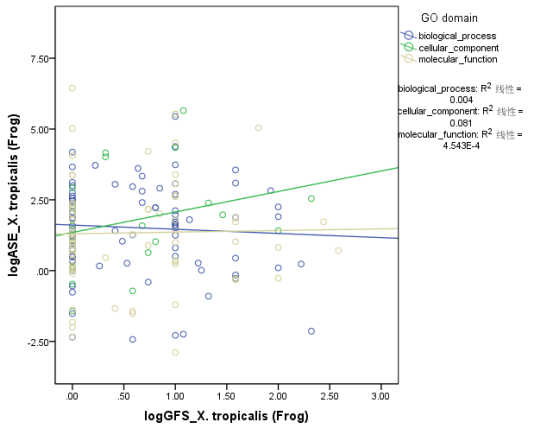


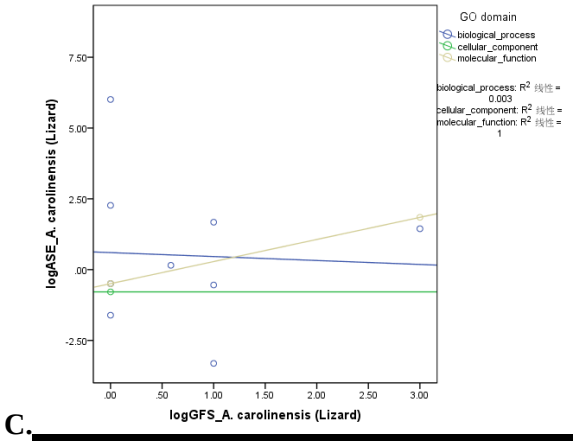
Figure S20 Co-relation between the log value of GFS and ASE change level from invertebrates of three genetic domains (biological process cellular component and molecular function) in different species (A) *D.rerio*, (B) *X. tropicalis* ,(C) *A. carolinensis*, (D) *G.gallus*, (E) *M.musculus*.



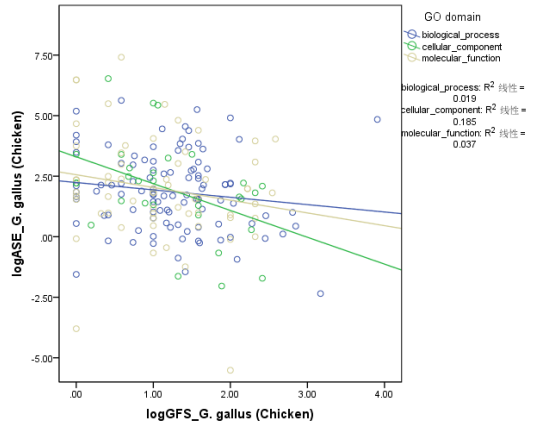
A.



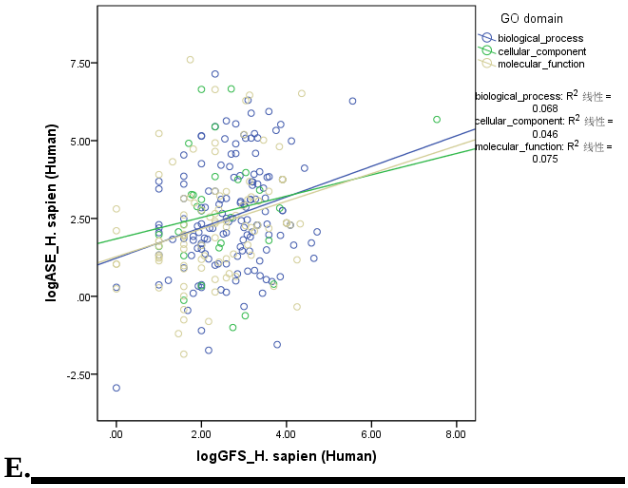
B.



C.



D.



E.

Figure S21 Co-relation between the log value of GFS and ASE change level from invertebrates singletons of three genetic domains (biological process cellular component and molecular function) in different species (A) *D.rerio*, (B) *X. tropicalis* ,(C) *A. carolinensis*, (D) *G.gallus*, (E) *H.sapien*.